Ein Verfahren zur Anreicherung fachgebietsspezifischer Ontologien durch Begriffsvorschläge

Am Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte Dissertationsschrift von

Dipl.-Math. Andreas Faatz

geboren am 30.11.1972 in Friedberg (Hessen)

zur Erlangung des Grades eines
Doktor-Ingenieurs (Dr.-Ing.)

Darmstadt 2004
Hochschulkennziffer D-17

Vorsitz:	Prof. Dr. Erich J. Neuhold
Erstreferent:	Prof. Dr.-Ing. Ralf Steinmetz
Korreferent:	Prof. Dr. Johannes Fürnkranz

Tag der Einreichung:	24.09.2004
Tag der Prüfung:	25.11.2004

Inhaltsverzeichnis

Kapitel 1

Einführung

Die Anwendung der Informations- und Kommunikationstechnologie führt zu einer explosionsartig wachsenden Menge potenziell nutzbarer Informationen. Für immer mehr Anwender stehen rechnergestützte Ansätze zur Automatisierung von Aufgaben und Tätigkeiten in zahlreichen Lebensbereichen zur Verfügung. Soll dieser Erfolg der Informations- und Kommunikationstechnologie fortgesetzt werden, ist eine noch engere Ausrichtung ihrer zukünftigen Entwicklung am menschlichen Wissen unabdingbar. Eine dauerhafte effiziente Orientierungsmöglichkeit innerhalb der Menge der Informationen kann geschaffen werden, wenn menschliches Wissen in Suchtechniken Eingang findet. Auch die Neuentwicklung von Algorithmen und das Zusammenspiel vorhandener Algorithmen, die eine weitere Automatisierung und Unterstützung menschlichen Handelns bewirken sollen, profitieren entscheidend von Techniken, die das menschliche Wissen verfügbar werden lassen. Eine solche Technik ist die Wissensrepräsentation.
Das generelle Ziel von Wissensrepräsentationen besteht in der formalen Darstellung eines Wissensgebietes. Das Ergebnis der Wissensrepräsentation ist eine Umsetzung der fachlich-thematischen Sicht derjenigen, welche die Wissensrepräsentation erschaffen. Diese Vertreter eines Wissensgebietes erzielen idealerweise einen fachlichen Konsens und stellen ihn formal dar. Dies ist ein anspruchsvoller und arbeitsintensiver Vorgang. Wenn Wissensrepräsentationen erfolgreich eingesetzt werden sollen, dann muss der Weg zur Erstellung einer Wissensrepräsentation so weit wie möglich erleichtert werden. Während zahlreiche Arbeiten, aktuell beispielsweise die um die Semantic Web Initiative positionierten Entwicklungen, die Nutzung einer rechnergestützten Wissensrepräsentation beschreiben und entwickeln, widmet sich ein vergleichsweise geringerer Teil der Arbeiten den expliziten Verfahren zur

Erstellung von Wissensrepäsentationen. Die vorliegende Dissertationsschrift legt den Schwerpunkt auf eine Erstellung durch eine Expertengruppe und deren Unterstützung durch den Rechner. Die Randbedingung dabei ist ein vertretbarer Aufwand bei der Erschaffung von Wissensrepräsentationen. Die Arbeit liefert einen Beitrag, der der automatischen Unterstützung des Erstellungsprozesses von Wissensrepräsentationen dient. Ziel ist ein Verfahren zur automatisch unterstützten Erstellung und Erweiterung eines bestimmten Typs von Wissensrepräsentationen, nämlich der so genannten Ontologien.
Für den Erstellungsprozess der Wissensrepräsentation selbst sollen wiederum die Vorteile eines intensiven methodischen Einsatzes der Informationstechnologie genutzt werden.
Aus dem dargestellten Zusammenhang ergeben sich Anforderungen, die den Gang der vorliegenden Arbeit und vordringlich die Definitionen in Kapitel 2 bestimmen werden:

- Jede Form der Wissensrepräsentation muss von ihrer Struktur und insbesondere von ihren strukturierenden Elementen her für ihre Ersteller und Nutzer klar verständlich sein. Dabei ist zu berücksichtigen, dass Ersteller und Nutzer in vielen Fällen Nichtinformatiker sind.

- Rechnergestützte Wissensrepräsentation muss einen Grad an Formalität aufweisen, der das menschliche Verständnis der Wissensrepräsentation gewährleistet. Wissensrepräsentation muss eine nachvollziehbare und umsetzbare Form annehmen und erklären, wie ihre Strukturelemente in Beziehung zueinander stehen.

Wäre eine von diesen Grundanforderungen verletzt, so könnte entweder die Mensch-Maschine-Interaktion für die Zwecke der Wissensrepäsentation nicht mehr effektiv genutzt werden, oder die Kommunikation und letztlich der Austausch des Wissens unterläge Hindernissen.
Vor dem Hintergrund eines möglichen Wissenswandels und Erkenntnisgewinns einerseits und der Wiederverwendung von Wissensrepräsentationen in einem neuen Anwendungszusammenhang andererseits kann als weitere Forderung gelten:

- Die Struktur der Wissensrepräsentation muss eine leichte Erweiterbarkeit gewährleisten.

Wir beschränken uns im Rahmen dieser Anforderungen auf die besondere Form der Wissensrepräsentation durch Ontologien. Ontologien müssen ein spezielles Wissensgebiet darstellen können. Den aus der Philosophie stammenden Totalanspruch auf Wirklichkeitsbetrachtung durch die Ontologie

(griechisch: 'Lehre von Seienden') lässt die Informatik fallen, arbeitet aber weiterhin mit **Begriffen** als strukturellen Einheiten der Wissensrepräsentation. Eine Ontologie besteht demnach aus Begriffen und Relationen, die die Begriffe sinnträchtig miteinander verbinden. Diese speziellere Art der Wissensrepräsentation führt zu einer Spezialisierung der Anforderungen. Wir übertragen die obigen allgemeinen Anforderungen an Wissensrepräsentationen auf Ontologien:

- Ontologien müssen von dem strukturellen Zusammenhang der Begriffe und Relationen her für ihre Ersteller und Nutzer, die auch Nichtinformatiker sein können, klar verständlich sein.
- Ontologien müssen einen Grad an Formalität aufweisen, der sowohl menschliches Verständnis als auch maschinelle Verarbeitung gewährleistet.
- Die Art, wie eine bestehende Ontologie zu erweitern ist, muss systematisch ihre bereits bestehenden Begriffe und Relationen berücksichtigen.

Als prinzipiellen Ansatz zur Erstellung und Erweiterung von Ontologien, welche die genannten Anforderungen erfüllen, stellt die vorliegende Arbeit auf die natürlichsprachliche Form eines Begriffes ab. Die natürliche Sprache soll als eine Quelle der Wissensrepräsentation fungieren, da natürliche Sprache einen allgemein zugänglichen Ausgangspunkt zur Formalisierung des menschlichen Wissens darstellt.
Das technische Ziel der vorliegenden Arbeit ist die Definition eines so genannten Ontologieanreicherungsverfahrens, das bei der Erstellung und Erweiterung einer gegebenen Ontologie neue Begriffe vorschlägt. Diese Begriffe sollen den bestehende Begriffen passend zugeordnet werden. Die letztendliche Akzeptanz und relationale Einbindung solcher Begriffsvorschläge obliegt den Erstellern der Ontologie. Das Verfahren kann in unserem Zusammenhang als erfolgreich betrachtet werden, wenn die Begriffsvorschläge eine Ontologie prinzipiell so erweitern, wie ein menschlicher Ersteller sie erweitern würde.

Aufbau der Arbeit

Wir benötigen eine Ontologiedefinition, die den obigen Anforderungen an Aufbau und Erweiterung von Ontologien Rechnung trägt und die (verschriftlichte) natürliche Sprache als Wissensquelle berücksichtigt. Kapitel 2 der

vorliegenden Arbeit greift eine solche Definition auf und entwickelt sie durch Zusatzanforderungen weiter. Zudem werden in Kapitel 2 die Einsatzmöglichkeiten von Ontologien, die der Definition entsprechen, umrissen.
Kapitel 3 richtet das Augenmerk auf den Erstellungs- und Erweiterungsaspekt. Ausgehend von der Prämisse, dass stets eine Arbeitsgruppe mehrerer Experten eines Wissensgebietes die Ontologie erstellt oder erweitert, wird ein wohl definierter, gruppenspezifischer Erstellungsprozess in einem konkreten Fall ausgearbeitet. Dabei werden wir sehen, dass - gerade hinsichtlich des Einsatzes natürlicher Sprache in Form von Fachtexten - die Möglichkeiten einer rechnerunterstützten Erweiterung bestehender Ontologien bislang nicht ausgeschöpft sind. Für eine automatische Unterstützung der Ontologieerstellung und -erweiterung leiten wir zu Ende des dritten Kapitels weitere Anforderungen ab. All diese beziehen sich im Kern darauf, dass die Ersteller der Ontologie eine weitere Wissensquelle in den Prozess einbringen können. Diese Quelle besteht aus fachgebietsbezogenen natürlichsprachlichen Texten. Des weiteren erklären wir, warum ein **Verständnis von Ähnlichkeiten zwischen Wörtern aus den fachgebietsspezifischen Texten** entscheidend zur Definition eines Ontologieanreicherungsverfahrens beitragen kann. In Folge dessen erstrecken sich die Untersuchungen verwandter Arbeiten in Kapitel 4 auf solche, die aus Textsammlungen automatisch Ontologien konstruieren. Wir werden zu Ende des vierten Kapitels eine noch zu schließende Lücke zwischen den Anforderungen aus Kapitel 3 und den verwandten Arbeiten aus Kapitel 4 identifizieren. Diese Lücke besteht in erster Linie im mangelnden Bezug verwandter Arbeiten auf die relativ geringen Restriktionen unserer Ontologiedefinition und in den spärlichen Erfahrungen mit Ansätzen zur Erweiterung fachspezifischer Ontologien. Wir grenzen uns zudem von der vollständig automatischen Konstruktion von Ontologien ab.
Kapitel 5 schließt die erwähnte Lücke durch die Entwicklung eines formalisierten Ontologienreicherungsansatzes. Außer der Existenz einer gegebenen Ontologie und einer fachspezifischen Textsammlung soll dabei möglichst wenig explizit formuliertes Zusatzwissen den Ansatz mitbestimmen. Die Ontologieanreicherung sieht Begriffsvorschläge als Erweiterung bestehender Ontologien vor. Ein Begriffsvorschlag entsteht dabei durch ein neuartiges Ähnlichkeitsverständnis für Wörter und Begriffe: erreicht ein Wort aus einem Textkorpus nach dieser Art des Vergleiches eine hohe Ähnlichkeit, so wird es zum Begriffsvorschlag. Das Ontologieanreicherungsverfahren wird für verschiedene mathematische Auffassungen von Ähnlichkeit formuliert. Zudem zeigen wir in Kapitel 5 eine automatische Evaluationsmethodik für Ontologieanreicherungen.
Kapitel 6 schließlich beschreibt die technische Umsetzung des Ontologiean-

reicherungsverfahrens und unterzieht sie einer Prüfung anhand der Gütekriterien aus Kapitel 5. Die Evaluation erweist sich für ein bestimmtes Szenario mit realistischen Datenbeständen als erfolgreich. Kapitel 6 schließt mit einem methodischen Vorschlag zur Anreicherung weiterer fachgebietsspezifischer Ontologien.
Die eigenen Beiträge des Verfassers erstrecken sich auf die genaue Spezifikation der kooperativen Ontologieerstellung in Kapitel 3, sowie auf die Entwicklung und Erpobung des Ontologieanreicherungsverfahrens in den Kapiteln 5 und 6. Der Hauptteil der Arbeit schließt mit einer Zusammenfassung der Ergebnisse.

Danksagung

Die Erstellung meiner Arbeit wurde wesentlich durch mein Umfeld ermöglicht und gefördert. Ich bin meinen Eltern und Großeltern für die geistige und materielle Unterstützung, die mich auf dem Weg durchs deutsche Bildungssystem begleitet hat, zutiefst dankbar. Ebenso möchte ich Herrn Prof. Dr.-Ing. Ralf Steinmetz für die Betreuung der vorliegenden Arbeit herzlich danken. Mein Dank gilt auch dem gesamten Lehrstuhl KOM, meiner Forschungsgruppe MMSem und meinem geduldigen Zimmergenossen Stefan Hörmann. Meinen besonderen Dank möchte ich all denen aussprechen, die die vorliegende Arbeit auf Fehler und Verbesserungswürdiges hin untersucht haben: Dr.-Ing. Cornelia Seeberg, Prof. Dr. Johannes Fürnkranz, Dr.-Ing. Michael Liepert, Dr. Andreas Mauthe, MA Annette Hansen und Assessor Jan Hansen. Meinen ontologiebegeisterten Studentinnen und Studenten Clara Möller, Melanie Dejardin, Alexander Elgert, Husam Qashgish, Ali Hayek und Lasse Lehmann möchte ich an dieser Stelle ebenfalls danken. Schließlich danke ich Carolin Kluge für den emotionalen Halt während der Vorbereitung und Ausführung dieser Arbeit.

Andreas Faatz, Darmstadt, September 2004

Kapitel 2

Ontologien

In diesem Kapitel werden die für die gesamte Arbeit grundlegenden Definitionen und Terminologien eingeführt. Neben einer Einordnung der Arbeit in die Aktivitäten des Forschungsgebietes Wissensrepräsentation sollen dabei hauptsächlich exakte Definitionen der Begriffe 'Ontologie' und 'Begriff' aus Sicht der Informatik entwickelt werden. Hierbei werden Vergleiche verschiedener gängiger Definitionen durchgeführt, und diese werden jeweils hinsichtlich der Problemstellung der gesamten Arbeit bewertet. Abschließend werden in der zweiten Hälfte des vorliegenden Kapitels die Ontologiedefinitionen um Anwendungsbeispiele ergänzt. Dabei wird aufgezeigt, welche Aspekte der vorher gefundenen Definitionen für diese Anwendungstypen benötigt werden.

2.1 Wissensrepräsentation

Die vorliegende Arbeit stellt ein Forschungsprogramm und Ergebnisse vor, die zum Gebiet der Wissensrepräsentation zu zählen sind. Es folgt daher eine kurze Einführung in die Fragestellungen der Wissensrepräsentation.
Das Forschungsgebiet Wissensrepräsentation beschäftigt sich mit methodisch fassbaren Abbildern von Ausschnitten der Wirklichkeit. Die übliche stufenweise Unterscheidung des informatischen Verständnisses der Wissensrepräsentation umfasst Zeichen, Daten, Information und Wissen [7].
Daten unterscheiden sich von Zeichen durch eine für sie geltende Syntax, Information fügt den Daten Bedeutung hinzu. Wissen schließlich wird bei dieser nach ihrem Sinngehalt abgestuften Unterscheidung als Information, die in Handlung umsetzbar ist, definiert [56]. Dies sei nun am Beispiel des Verlaufes von Börsenindices verdeutlicht. Der DAX beispielsweise stellt den

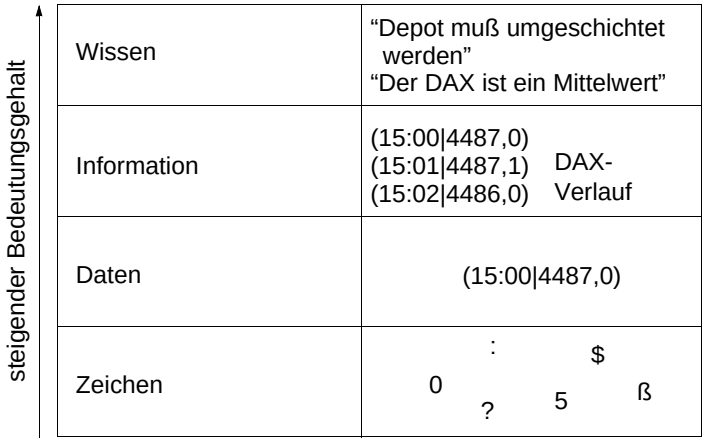

Abbildung 2.1: Stufenweise Unterscheidung des Bedeutungsgehaltes

gemittelten Wert ausgewählter deutscher Aktien im Zeitverlauf dar. Das Verhältnis des Wissens um den DAX zu konkreten Daten lässt sich wie folgt erklären. Die Abbildung 2.1 zeigt auf der untersten Bedeutungsebene Zeichen wie '5', '0' und ':'. Paare von Zeichenketten wie '15:00' und '4487,0' bilden die Grundelemente eines numerisch aufgezeichneten Verlaufes des Börsenindex DAX. Syntaktisch soll in unserem Beispiel gelten, dass der erste Teil eines solchen Paares wie (15:00—4487,0) von zwei durch einen Doppelpunkt getrennte zweistellige Zahlen gebildet wird. Die erste nimmt Werte zwischen '00' und '23', die zweite nimmt Werte zwischen '00' und '59' an, wobei wir sehen, dass einstellige Dezimalzahlen jeweils als zweistellig dargestellt werden. Der zweite Teil des Paares von Zeichenketten besteht aus einer Dezimalzahl mit einer Nachkommastelle. Des Weiteren seien nun viele solcher Paare so aufgelistet, dass die erste Zeichenkette sie ordnet:

...
(15:00—4487,0)
(15:01—4487,1)
(15:02—4486,0)
(15:03—4486,9)

(15:04—4487,0)
(15:05—4487,1)
(15:06—4486,7)
...

Eine solche Ordnung und Auflistung der Paare stellt, wenn der erste Teil der Paare alle Werte von (09:00) bis (17:00) durchläuft, die Art dar, wie Daten über den Verlauf eines Börsenindex beschaffen sein müssen. In den folgenden Überlegungen gehen wir davon aus, dass die Daten korrekt erfasst und aufgezeichnet wurden.
Nach [65] sind Daten in einem Kommunikationsablauf sprecherunabhängig und hörerunabhängig, das heißt, es muss sichergestellt sein, dass alle Sprecher (in unserem Falle diejenigen, die die Auflistung der Paare erstellt haben) die gleichen syntaktischen Regeln einhalten, und dass dadurch alle Hörer (in unserem Fall diejenigen, die die Auflistung durchlesen oder allgemeiner: erfassen und gegebenenfalls weiterverarbeiten) die gleichen Voraussetzungen einer Deutung der Daten haben. Die in den Daten erhaltene Information kann hingegen durchaus hörerabhängig sein. Ein Wirtschaftswissenschaftler wird in der Lage sein, eine exakte Definition für das Zustandekommen des zweiten Eintrages in einem Datenpaar (Berechnung des Börsenindex) zu liefern, während ein Privatanleger diese Definition nicht genau kennen muss, trotzdem aber den Daten Bedeutung zukommen lässt, indem er den ersten Eintrag als Uhrzeit, den zweiten Eintrag als Stand des Börsenindex und die Abfolge der Paare als zeitlichen Verlauf des Index interpretiert.
Bei der Rolle des Wissens gehen nun die Auffassungen der Literatur auseinander. Während es eine Richtung gibt, die Wissen in weiten Teilen als durch Information bedingt und erzeugt ansieht, gilt für Vertreter des Konstruktivismus [77], dass das Wissen erst die Bedeutung zu den Daten hinzufügt. Im Falle des obigen Beispiels wären also das Wissen des Wirtschaftswissenschaftlers und das Wissen des Privatanlegers ausschlaggebend für die Herstellung von Information. Unabhängig davon, ob wir diese konstruktivistische Sicht übernehmen oder nicht, halten wir fest, dass Wissen weit über Information hinausgeht. Der Wirtschaftswissenschaftler wird die Information zum DAX-Verlauf in einen theoretischen Kontext einzubinden wissen, in Bezug setzen zu Theorien und Erfahrungen, die mit Begriffen wie 'Zyklen', 'Konjunktur', 'Chartanalyse' oder 'Verbraucherschutz' zusammenhängen. Das handlungsbezogene Wissen des Privatanlegers kann sich bei bestimmten Verläufen des DAX darin äußern, zum Telefonhörer zu greifen und ein Beratungsgespräch mit der persönlichen Betreuerin bei der Bank seines Vertrauens zu führen. Beiden gemein ist ein Operieren mit Hilfe von

Bezügen zwischen den ihr Wissen strukturierenden Einheiten. Im ersten Fall könnte ein Bezug 'Entwicklungen der Konjunktur verlaufen in Zyklen' identifiziert werden, im zweiten Fall Bezüge wie 'die Bank XYZ besitzt ein Beratungsangebot' oder - was ein wesentlich unspezifischeres, wenig auf die Wirtschaftswissenschaften an sich bezogenes Wissen darstellt - 'zum Herstellen einer Telefonverbindung benötigt man ein Telefon und eine Telefonnummer'. Wissen ist somit in verschiedene Ausprägungen klassifizierbar. Eine Klassifikation nach Anzahl der das Wissen Kommunizierenden umfasst

- persönliches oder subjektives Wissen, das hauptsächlich einem Handelnden zur Verfügung steht. In unserem obigen Beispiel kann 'die Bank XYZ ist die Bank meines Vertrauens' als das subjektive Wissen des Privatanlegers gelten. Das Denken als Operation mit dieser Form des Wissens wird bei Luft et al. [65] als Kommunikation eines Subjektes mit sich selbst verstanden.

- organisationales Wissen, das innerhalb einer Organisation von mehreren geteilt wird. Hier liegt eine Form von Wissen vor, die einer begrenzten Anzahl von Personen zur Verfügung steht. Innerhalb einer Geschäftsbank kann dieses Wissen beispielsweise die dort bekannte Tatsache umfassen, dass Mitarbeiterin B. eine hervorragende Spezialistin für risikoreiche Aktiendepots ist und daher bestimmte Kunden beraten sollte.

- objektives oder objektivierbares Wissen, das mit von einer großen Gemeinschaft anerkannten und überprüfbaren Methoden geschaffen und erweitert wird. Die Kenntnisse des Wirtschaftswissenschaftlers sind hierunter zu fassen.

Diese Einteilung verläuft zunächst unabhängig davon, ob die jeweilige Art von Wissen formalisierbar ist oder nicht. Eine weitere Einteilung, die ebenfalls bereits aus obigem Beispiel der Unterscheidung zwischen Wissen, Information und Daten hervorgeht, ist die nach dem Spezialisierungsgrad des Wissens:

- Allgemeinwissen umfasst den Zusammenhang alltäglicher Handlungen.

- fachgebietsspezifisches Wissen besitzt einen begrenzten Gültigkeitsbereich.

Die vorliegende Arbeit befasst sich mit Methoden, die für die speziellere Form von Wissen relevant sind und geht gleichzeitig davon aus, dass mehrere Personen am systematischen Aufbau eines Wissensbestands beteiligt

sind.
Das Forschungsgebiet Wissensrepräsentation operiert auf zwei Weisen, die allerdings nicht strikt von einander zu trennen sind. Die erste Betrachtungsweise stellt die Frage, wie das menschliche Bewusstsein Wissen erfasst, erzeugt, darstellt und dauerhaft verarbeitet, während die zweite versucht, Wissen methodisch zu formalisieren. In beiden Bereichen ist das Wirken der Sprache und die Wechselwirkung zwischen Sprache und Wissen einer der zentralen Gegenstände der Untersuchungen. Die vorliegende Arbeit steht in ihren Ansprüchen und Methoden jedoch im Zusammenhang mit der zweiten Lesart der Wissensrepräsentation. Zur ersten sei an dieser Stelle lediglich erwähnt, dass es stark voneinander abweichende Standpunkte zu der Frage gibt, ob und wie das menschliche Bewusstsein ein Abbild der Wirklichkeit herstellt. Durch Arbeiten zur biologischen Systemtheorie [70] beispielsweise wurde diese Frage für die biologische Analyse von Erkenntnisprozessen als irrelevant erklärt, während andere Arbeiten von begrifflichen Repräsentationen oder Manifestationen in kognitiven Strukturen ausgehen [81].
Als Grundelemente der Wissensrepräsentation dienen in unserem Zusammenhang Begriffe, die als charakteristische sinnträchtige Einheiten eines Wissensbereiches verstanden werden können. Unsere Forschungsmethodik wird in einer Weise vorgehen, die über eine schiere Aufzählung vorhandenen Wissens, wie sie beispielsweise durch eine lexikalische Auflistung vorhandener Fachbegriffe aus einem bestimmten Wissensgebiet entstehen kann, wesentlich hinausgeht: sie wird sich mit Strukturen beschäftigen, die die Begriffe als Bestandteile fachgebietsspezifischen Wissens nicht nur identifizieren, sondern auch einen inhaltlichen Zusammenhang zwischen den einzelnen Wissenskomponenten definieren. Dieser inhaltliche Zusammenhang soll als Grundlage dazu dienen, weitere, noch nicht erfasste Zusammenhänge zu zusätzlichen Begriffen zu erkennen.
Eine Formalisierung von Wissen identifiziert im Allgemeinen die Elemente des Wissensgebietes in Form von Symbolen, Variablen und Formeln, wie dies beispielsweise in der Prädikatenlogik geschieht. Im Speziellen wird eine solche Formalisierung mittels einer Programmiersprache elektronisch repräsentiert, wie dies beispielsweise in der Programmiersprache Prolog für die Prädikatenlogik der Fall ist. Aus diesem Zusammenhang heraus stellt sich die Frage, wie formalisiertes Wissen maschinell erfasst, dargestellt und verarbeitet werden kann. Als zusätzliche Annahme soll hierbei gelten, dass der Verarbeitung des formalisierten Wissens auch der Zugriff auf in natürlicher Sprache verfasste Quellen erlaubt ist.
Die Art der hier zu untersuchenden Verarbeitung liegt vor allem in der Erweiterung vorhandener Wissensbestände. Im Zuge der vorliegenden Arbeit

wird somit die Frage untersucht, inwieweit und auf welche Weise natürliche Sprache in Form großer Textmengen eine zusätzliche geeignete Quelle zum Aufbau formaler Wissensrepräsentationen bilden kann. Da wir Verfahren zur Erweiterung der durch formale Wissensrepräsentation erfassten Wissensbestände untersuchen, wird im Sinne von [71] auch von 'Lernen' oder 'maschinellem Lernen' die Rede sein.
Dabei geht es uns vordringlich um die Darstellung speziellen, fachgebietsspezifischen Wissens, das zunächst nur ein eingeschränktes, möglichst genau zu definierendes Wissensgebiet betrifft. Die Darstellung des allgemeinen Wissens oder auch Weltwissens, das wir benötigen, um alltägliche Tätigkeiten und Situationen verrichten und bewältigen zu können, ist zwar ebenfalls ein anspruchsvoller Forschungsgegenstand der Wissensrepräsentation, wird jedoch nicht im Blickpunkt dieser Arbeit stehen.
Als Anforderungen für den Gang der Arbeit lassen sich nun mehrere Anforderungen an die Wissensrepräsentation nennen, die allesamt ihren Praxisbezug gewährleisten sollen und sie von rein philosophischen Betrachtungen etwa im Rahmen der Erkenntnistheorie oder von psychologischen Betrachtungen etwa im Rahmen der Kognitionsforschung unterscheiden sollen:

- Wissensrepräsentation soll fachgebietsspezifisch sein.
- Wissensrepräsentation soll mit einem moderaten Schulungsaufwand vermittelbar und erstellbar sein.
- Wissensrepräsentation soll wiederverwendbar sein, um Ressourcen schonend eingesetzt werden zu können.
- Wissensrepräsentation soll prinzipiell in einer Form erfolgen, die sowohl für Menschen als auch für Maschinen lesbar ist.

In den folgenden Abschnitten wird der Bezug zu fachgebietsspezifischem Wissen vor allem durch beispielhafte Anwendungen und anhand des Erstellungsprozesses formaler Wissensrepräsentationen in Form von Ontologien aufgezeigt. Ontologien werden in einer Form eingeführt, die Teile der Anforderungen an die Wissensrepräsentation bereits gut erfüllen. Die gesamte Arbeit wird den Versuch darstellen, eine weitere Erleichterung bei der Erstellung von Wissensrepräsentationen durch Ontologien zu liefern und somit eine weitere Annäherung an die obigen Anforderungen zu erreichen.

2.2 Begriffe und Ontologien

Der nun anschließende Abschnitt dient der Herausarbeitung einer Ontologiedefinition. Die in der Definition herausgearbeitete Struktur wird im Laufe der Arbeit sowohl für die Entwicklung eines Ontologieanreicherungsverfahrens, als auch für dessen Bewertung verwendet.
Wie in der Einleitung gefordert, benötigen wir eine Ontologiedefinition, die von einer Notation der Ontologie mittels einer konkreten formalen Sprache unabhängig ist, um einen Zugang für Nichtinformatiker zu gewährleisten. Gleichzeitig soll die Formalität der Definition als Grundlage für Ontologieerstellungsprozesse und insbesondere für Ontologieanreicherungen durch Begriffsvorschläge dienen.

2.2.1 Begriffe und ihre Ordnung

Eingangs wurde für nutzerorientierte Formen der Wissensrepräsentation gefordert, dass ihre Erweiterung aus ihrer Struktur heraus erklärbar sein muss. Der vorliegende Abschnitt geht auf Begriffe und die sie ordnende Unterbegriffsbeziehung ein, die beide als wesentliche Bestandteile in die Ontologiedefinition, in den Ontologieanreicherungsalgorithmus und dessen Bewertung eingehen werden.
Der folgende Abschnitt führt Notationen und Begriffskonstruktionen der formalen Begriffsanalyse ein. Zum einen wird durch diesen Formalismus in kompakter Darstellungsform deutlich, welche Ordnungseigenschaften Begriffe in der Wissensrepräsentation haben müssen. Zum anderen wird im Kapitel der verwandten Ansätze der vorliegenden Arbeit ein Rückgriff auf die Formalismen des folgenden Abschnittes vorgenommen.
Begriffe wurden in aktuelleren Arbeiten (unter anderem von Seiler [90]) als Grundeinheiten menschlichen Denkens beschrieben, denen eine zentrale Rolle im Denkprozess zukommt. Begriffe können intensional oder extensional definiert sein. Die extensionale Definition eines Begriffes stellt eine gedankliche Einheit eines oder mehrerer Gegenstände her. Beispielsweise könnte man die Mitarbeiter Schulze, Maier und Schmidt als Teilmenge aller Mitarbeiter einer Firma H. gedanklich unter dem Begriff 'verdiente Mitarbeiter der Firma H.' zusammenfassen, wenn diese Zusammenfassung jemandem als wahrheitsgemäß und sinnvoll erscheint. Über die konkreten Eigenschaften dieser Mitarbeiter (Begriffsintension) ist damit noch nichts gesagt. Umgekehrt könnte der Begriff 'verdiente Mitarbeiter der Firma H.' aber auch über Eigenschaften von Mitarbeitern definiert werden, womit allerdings noch nicht die Menge der Mitarbeiter (Begriffsextension) ausformuliert ist, die

diese Eigenschaften erfüllen. Eine Konstruktion von Begriffen aus einer so genannten dyadischen Sicht, die die Verstrickung von Begriffsintension und Begriffsextension berücksichtigt, liefert die von Rudolf Wille entwickelte formale Begriffsanalyse. Hier wird die Entwicklung von Begriffen aus vorliegenden Gegenständen und ihren feststellbaren Eigenschaften sichtbar. Die Abbildung 2.2 zeigt in tabellarischer Form ein Beispiel für einen so genannten formalen Kontext nach Wille. Pro Zeile sei hier ein Mitarbeiter aufgetragen, das Vorhandensein eines Kreuzchens bedeutet dann, dass die in der entsprechenden Spalte aufgeführte Eigenschaft erfüllt ist. Eine algebraische Schreibweise für einen formalen Kontext liefert die folgende Definition nach Wille [104]:

Definition 1 (formaler Kontext) *Seien G und M Mengen und I eine Relation zwischen G und M, das heißt $I \subseteq (G \times M)$. Ein $formalerKontext$ ist eine Mengenstruktur $K := (G, M, I)$. Die Elemente von G nennt man Gegenstände, die Elemente von M Merkmale und gIm steht dafür, dass der Gegenstand g das Merkmal m besitzt. Die folgenden Ableitungsoperatoren sind für alle $X \subseteq G$ und für alle $Y \subseteq M$ definiert:*

$$X \mapsto X^I := \{m \in M | gIm \text{ für alle } g \in X\} \ ,$$

$$Y \mapsto Y^I := \{g \in G | gIm \text{ für alle } m \in Y\}.$$

Wir stellen fest, dass diese allgemeine Definition eines formalen Kontexts nicht festlegt, welche Mengen von Gegenständen und Merkmalen betrachtet werden. Vielmehr kann aus jeder Menge eine Gegenstandsmenge und aus jeder Menge eine Merkmalsmenge entstehen, wenn die Relation I existiert. Obwohl wir in der zu entwickelnden Ontologiedefinition von den *konkreten* Merkmalen und Gegenständen absehen, eignet sich die Herleitung von Begriffen aus formalen Kontexten, um auch für Ontologien gültige begriffliche Strukturen sichtbar werden zu lassen.
Für die Ableitungsoperationen beschreiben die folgenden Gleichungen das Enthaltensein von Gegenstandsmengen und Merkmalsmengen (Z kann hier für beides stehen) , nämlich eine entgegengesetzte Beziehung für Merkmale und Gegenstände

$$Z_1 \subseteq Z_2 \Rightarrow {Z_1}^I \supseteq {Z_2}^I, \tag{2.1}$$

und das Enthaltensein unter zweifacher Ableitung

$$Z \subseteq Z^{II}, \tag{2.2}$$

	Englisch verhandlungs-sicher	*kennt asiatische Märkte*	*kennt Kredit-vergabe*	*kann gut erklären*	*länger als 3 Jahre angestellt*	*kürzer als 3 Jahre angestellt*
Becker		X	X			X
Müller	X		X			X
Maier	X		X			X
Lehmann	X		X	X	X	
Schmidt		X		X	X	
Schulze		X	X		X	

Abbildung 2.2: Formaler Kontext

mit $Z^{II} := (Z^I)^I$. Zudem gilt die Identität

$$Z^{III} = Z^I, \tag{2.3}$$

welche zeigt, dass die wiederholte Anwendung der Ableitungsoperation schließlich einen stabilen Zustand herstellt und nicht bei ihrer Anwendung die jeweilige Merkmals- oder Gegenstandsmenge beliebig vergrößert oder verkleinert. Die folgende Definition formaler Begriffe nach Wille [104] nutzt diese Eigenschaften aus:

Definition 2 (formaler Begriff) *ein formaler Begriff eines formalen Kontexts $K := (G, M, I)$ ist ein Paar (A, B) mit $A \subseteq G$, $B \subseteq M$, $A = B^I$ und $B = A^I$. A wird die Extension, B die Intension eines formalen Begriffes genannt.*

Formale Begriffe stehen in einer Ober-Unterbegriffsrelation zueinander.

Definition 3 (formale Unterbegriffsrelation) *die Relation $\leq$ zwischen zwei formalen Begriffen (A_1, B_1) und (A_2, B_2), welche durch*

$$(A_1, B_1) \leq (A_2, B_2) :\Leftrightarrow ((A_1 \subseteq A_2) \tag{2.4}$$

gegeben ist, heißt 'ist Unterbegriff von'.

Aufgrund der sie definierenden Mengeninklusionen sehen wir, dass die Ober-Unterbegriffsrelation eine Ordnungsrelation ist, das heißt sie ist reflexiv, transitiv und asymmetrisch. Daraus folgt auch, dass sie azyklisch ist. Die Definition der formalen Unterbegriffsrelation ist sowohl über die Gegenstände als auch über die Merkmale formulierbar, da

$$((A_1 \subseteq A_2) \Leftrightarrow (B_2 \supseteq B_1)) \tag{2.5}$$

gilt. Die obigen Eigenschaften lassen sich wie folgt notieren. Sei X eine Menge und $\sim$ eine Relation auf X

- $\forall x \in X : x \sim x$ (Reflexivität von $\sim$)
- $\forall x, y \in X : (x \sim y \wedge y \sim z) \Rightarrow (x \sim z)$ (Transitivität von $\sim$)
- $\forall x, y \in X : (x \sim y \wedge y \sim x) \Rightarrow x = y$ (Asymmetrie von $\sim$)

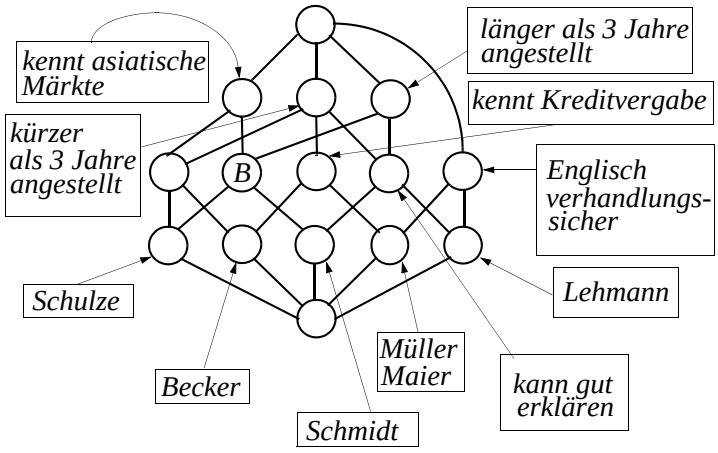

Abbildung 2.3: Begriffsverband als Liniendiagramm

Diese Eigenschaften gelten für die Relation $\leq$ ('ist Unterbegriff von') und wir werden sie für die Unterbegriffsrelation in Ontologien ebenfalls fordern. Im Folgenden wird davon ausgegangen, dass alle Begriffe (wie etwa 'guter Mitarbeiter') nicht ohne die Existenz bestimmter Eigenschaften existieren können. Dabei muss es sich nicht um die im Beispiel genannten handeln. Die vorliegende Arbeit teilt insofern die Auffassung Wittgensteins in dessen Bemerkung, dass 'ein Fleck nicht ohne seine Farbe (...), ein Ton nicht ohne seine Höhe' denkbar sei[107]. Wir werden davon ausgehen, dass Begriffe im Allgemeinen nicht ohne ihre Intensionen gedacht werden können. Im Allgemeinen jedoch wird die vollständige Konstruktion eines *konkreten* formalen Kontexts und der resultierenden formalen Begriffe ausbleiben. Eine Ober-Unterbegriffsrelation muss aber stets eine Ordnungsrelation sein, solange wir annehmen, dass ein formaler Kontext *existiert*. Dies rechtfertigt die Übertragung der formalen Begriffsanalyse auf die Ober-Unterbegriffsrelationen in Ontologien.
Abschließend weisen wir auf die graphische Darstellung formaler Begriffe im Liniendiagramm in Abbildung 2.3 hin. Jeder Knoten stellt einen formalen Begriff dar, die Beschriftungen Merkmale beziehungsweise Gegenstände.

Liest man das Diagramm entlang der Kanten von unten nach oben, so gilt zwischen Anfangs- und Endbegriff die Relation 'ist Unterbegriff von'. Die Gesamtheit der Begriffe der Ebene unterhalb des obersten Begriffes im Diagramm bildet dabei die Extension, die Gesamtheit der Begriffe der Ebene oberhalb des untersten Begriffes die Intension eines Begriffes.
Zusätzlich zu den gezeigten Eigenschaften der Ober-Unterbegriffsrelation gehen wir in dieser Arbeit mit [73] davon aus, dass alle Begriffe einen natürlichsprachlichen Namen besitzen. Mit Hilfe dieses Namens kann auf der Begriff als Einheit referenziert werden. In unserer bisherigen Überlegung ist ein solcher Name nicht gegeben, er ist auch nicht im Liniendiagramm 2.3 aufgetragen: dort existiert beispielsweise für den Begriff B mit der Intension {kennt asiatische Märkte, länger als 3 Jahre angestellt } kein Name wie 'Asienexperte' oder ähnliches. Mudersbach[73] fordert einen Formalismus für Begriffe, der ihrem Sprachgebrauch nahe kommt. Dieser Forderung liegt eine Betrachtung zugrunde, die die Prädikatenlogik in dieser Frage als nur eingeschränkt funktionstüchtig erklärt. Eine Unterscheidung der folgenden gesetzesartigen Aussagen soll demnach nicht mehr getroffen werden:

(a) alle Asienexperten sind Volkswirte
(b) der Asienexperte ist immer Volkswirt
(c) eine Asienexpertin ist eine Volkswirtin
(d) Asienexperten sind Volkswirte
(e) wenn jemand Asienexperte ist, dann ist er oder sie auch Volkswirt

Vielmehr soll ein Formalismus, der Begriffe als zugrunde liegende Einheiten verwendet, die Aussagen (a) bis (e) als identisch betrachten:

'Asienexperte'-'ist Unterbegriff von'-'Volkswirt'

Folgende Betrachtung des vorliegenden Abschnittes bleibt als auf Ontologien zu übertragende Grundannahme fest zu halten:

- es existiert eine Ordnungsrelation (Ober-Unterbegriffsrelation) der Begriffe eines Wissensgebietes

Formale Begriffsanalyse liefert dabei in einsichtiger Weise intensionale, extensionale und dyadische [105], das heißt gleichzeitig intensionale und extensionale, Begründungen der Transitivität, Reflexivität und Asymmetrie der Unterbegriffsrelation. Die formalen Begriffe tragen nicht notwendigerweise natürlichsprachliche Namen, wie etwa im Falle des in diesem Sinne namenlosen formalen Begriffes B in Abbildung 2.3. Allerdings verlangt die praxisorientierte Wissensrepräsentation, wie sie eingangs gefordert wurde, Begriffe

nicht als formale Begriffe (Knoten des aus dem formalen Kontext generierten Liniendiagramms). Vielmehr sollen Begriffe als gedankliche Einheiten sprachlich fixiert werden können und dabei sehr wohl die hier vorgestellte Ordnungsrelation analog zu den formalen Begriffen einhalten. Wir fordern:

- Begriffe besitzen einen natürlichsprachlichen Namen

Es werden nun verschiedene Definitionen der begrifflichen Wissensrepräsentation auf eine Einhaltung unserer Anforderungen hin zu untersuchen sein. Dabei beginnen wir mit semantischen Netzen sowie Topic Maps und setzen die Untersuchung mit Ontologien fort.

2.2.2 Semantische Netze

Eine von der Prädikatenlogik abweichende Form der Wissensrepräsentation ist die der Semantischen Netze. 'Semantisches Netz' deutet darauf hin, dass seine Struktur in einer Visualisierung ihre Entsprechung findet.
Nach Scragg [88] ist ein Semantisches Netz eine Ansammlung von Knoten, die als Konzepte interpretiert werden[1]. Jeder Knoten besitzt einen Namen, der aus einer Zeichenkette besteht. Knoten werden durch gerichtete Kanten, die die Relationen darstellen, verbunden. Besteht eine Relation in eine Richtung, dann gilt sie im Allgemeinen nicht auch in der umgekehrten Richtung. Die kleinsten Informationseinheiten in einem Semantischen Netz sind zwei durch eine Relation miteinander verbundene Knoten. Ein Konzept enthält nur Informationen, wenn es mit einem anderen Konzept in Relation steht, ansonsten ist das Konzept leer. Bei den Konzepten eines Semantischen Netzes wird zwischen Token und Typen unterschieden. Ein Token ist ein spezifisches Konzept wie das 'Bankkonto', mit der Bankleitzahl 508 501 50 und der Kontonummer 0100676532, Typ hingegen ist nicht spezifisch - wie beispielsweise das Konzept 'Bankkonto'. 'Bankkonto' ist über eine 'istEin'-Relation

[1] An dieser Stelle ist eine Vorbemerkung zum Sprachgebrauch des vorliegenden und des nächsten Abschnittes nötig. Während die Elemente Semantischer Netze sowohl in der englischsprachigen als auch in der deutschsprachigen Literatur als Konzepte (beziehungsweise 'concepts') bezeichnet werden, ist die korrekte Übersetzung des englischen 'concept' im Zusammenhang mit Ontologien 'Begriff'. Die Situation stellt sich derart dar, dass es im Deutschen zwei Bezeichnungen für zwei verschiedene Dinge gibt, dass das das englische 'concept' somit je nach Zusammenhang verschieden interpretiert werden muss. Die Hintergründe dieser Sprachunterschiede sind, wie wir sehen werden, inhaltlicher Art, und wir weisen nochmals ausdrücklich darauf hin, dass die Verwendung des deutschen Wortes 'Konzept' als Bestandteil von Ontologien falsch, als Bestandteil von Semantischen Netzen aber richtig ist.

mit 'Finanzinstrument' verbunden und ist somit ein Token vom Typ 'Finanzinstrument'. Der Unterschied zwischen Token und Typ wird in einem Semantischen Netz mittels verschieden benannter Relationen dargestellt. Um Token und Typ miteinander zu verbinden, könnte die 'istEin'-Relation verwendet werden, und für Relationen zwischen Typen die Subtyp-Relation. Dann besagt die Verkettung

Konto Nummer 347212 'istEin' Bankkonto 'Subtyp' Finanzinstrument,

dass das Konto mit der betreffenden Nummer ein Token vom Typ 'Bankkonto' und 'Bankkonto' ein Untertyp des Typs 'Finanzinstrument' ist. Die Struktur solcher 'Subtyp'-Relationen, ihre Entsprechung zu den Begriffsordnungen, die wir im letzten Abschnitt eingeführt haben und ihre Notwendigkeit beim Aufbau eines semantischen Netzes *an sich* werden allerdings in der Literatur nicht deutlich erklärt. Im Fall der von Sowa [94] eingeführten Conceptual Graphs führt dies dazu, dass Sowa zusätzlich den Begriff Ontologie einführt und besagte Begriffsstruktur teilweise auslagert. Andere Definitionen Semantischer Netze treffen keine explizite Annahme über das Vorhandensein einer solchen Struktur. Als Beispiel hierfür sind Topic Maps zu nennen. Topic Maps [108] sind ein ISO-Standard, der es ausdrücklich erlaubt, neben den obigen Arten von Konzepten auch Referenzen auf Dokumente in das durch sie repräsentierte Semantische Netz mit aufzunehmen. Die Verbindungen zwischen den Konzepten (Sprachgebrauch hier: Topics) einer Topic Map werden selbst wieder als Konzepte (Topics) betrachtet. Eine 'vomTyp'-Relation kann innerhalb einer Topic Map erklärt werden, ist jedoch nicht zwingend vorhanden. Eine Besonderheit der Topic Maps ist die mögliche Angabe des so genannten Scopes, der einen Gültigkeitsbereich für das repräsentierte Wissen angibt. Diese Scopes existieren innerhalb einer Topic Map, somit kann eine Topic Map auch Wissen über mehrere Gültigkeitsbereiche hinweg repräsentieren und Bedeutungen unterscheiden. Dies ist insofern relevant, als dass die Konzepte (Topics) einer Topic Map in natürlicher Sprache vorliegen. Im Scope 'Online-Auktionen' wäre 'Konto' somit ein elektronisch verfügbares Konto, im Scope 'Bank' besitzt 'Konto' eine allgemeinere Bedeutung. Darüber hinaus kann der Sinngehalt von Homonymen über verschiedene Scopes unterschieden werden.
Durch Ableitungsprozeduren können Inferenzen in einem Semantischen Netz verwirklicht werden. Beispielsweise könnte zwischen den Konzepten 'Konto' und 'Bankkunde' die Relation 'besitzt' bestehen. Dann gälte auch für ein Konzept 'Herr Tischbein', wenn es als Token vom Typ 'Bankkunde' vorläge, die Relation 'besitzt' zu 'Bankkonto'. Allerdings muss diese Verbindung nicht noch einmal getrennt aufgeführt werden, da sie über das Kon-

zept 'Bankkunde' schon hergeleitet werden kann. An dieser Stelle wird eine Analogie zu Vererbungsmechanismen der objektorientierten Programmierung sichtbar[75].
Wir werden im nächsten Abschnitt zeigen, dass alle gängigen Definitionen von Ontologien in der Informatik als Spezialfälle von Semantischen Netzen aufgefasst werden können, wir indes eine bestimmte Definition von Ontologien den anderen vorziehen werden, da sie sehr deutlich die im vorherigen Abschnitt betrachteten Eigenschaften begrifflichen Wissens umfasst.

2.2.3 Ontologiedefinitionen

Im Bezug auf das Gesamtziel der Arbeit gilt es nun, eine Ontologiedefinition zu finden, die den in der Einleitung geforderten Umgang mit Wissensrepräsentationen ermöglicht, und die als Grundlage für automatische Verfahren zur Erweiterung einer Ontologie dienen kann.
Das Wort Ontologie stammt aus dem Griechischen und bedeutet ' die Lehre vom Sein, von den Ordnungs-, Begriffs- und Wesensbestimmungen des Seienden' [22] . Die Verwendung des Wortes Ontologie bezeichnet somit ursprünglich eine Disziplin der Philosophie. In der vorliegenden Arbeit werden wir Ontologien nach einem wesentlich engeren Verständnis verwenden, nämlich dem der Informatik. Eingangs wenden wir uns einem linguistischen Verständnis von Ontologien zu, das auf die Definitionen der Informatik hinführt.
Nach Witmer [106] sind Ontologien aus philosophisch-linguistischer Sicht der Versuch zu bestimmen, was existiert. Aufgrund der Uneindeutigkeit der natürlichen Sprache fallen diese Entscheidungen mit den Mitteln natürlicher Sprache schwer. Witmer [106] versucht die natürliche Sprache einer Analyse hinsichtlich der ontologischen Bestimmungen von Begriffen zu unterziehen. Hat eine Sache einen Namen oder beinhaltet oder impliziert ein Satz eine existenzielle Generalisierung, für die die Sache vorhanden sein muss, dann existiert die Sache (entweder als gedankliches Konstrukt oder real) und hat eine Bestimmung in der Ontologie. Für den ersten Fall (explizite Erwähnung) sei der Satz

' Herr Tischbein bezahlt per Kreditkarte'

ein Beispiel. Der Satz deutet eindeutig darauf hin, dass die Person des Herrn Tischbein und eine Kreditkarte existieren. Für den zweiten Fall (Implikationen des Gesagten) sei der Satz

'Herr Tischbein überzieht sein Konto'

als Beispiel gegeben. Die Tatsache, dass Herr Tischbein sein Konto überzieht, legt nahe, dass er dazu berechtigt ist, somit sollte in der Ontologie der Begriff Dispo oder Kredit existieren, obwohl beides in keinem natürlichsprachlichen Ausdruck explizite Erwähnung findet.
Zu der im Beispiel illustrierten Schwierigkeit, aus natürlichssprachlichen Aussagen auf die Existenz der für eine Ontologie wesentlichen Begriffe zu schließen tritt für die Wissensrepräsentation eine weitere konstituierende Fragestellung hinzu. Gemäß des im vorigen Abschnitt erläuterten Verständnisses von Wissen muss eine Ontologie als Struktur der formalen Wissensrepräsentation Bezüge zwischen ihren Bestandteilen, die die Existenz von Dingen fassen, herstellen. Wäre dies nicht der Fall, so wäre es lediglich gerechtfertigt, die vorliegende Struktur als Information, jedoch nicht als Wissen zu bezeichnen.
Je nachdem, wie stark diese Forderung gewichtet wird und wie reichhaltig diese Bezüge zwischen den Bestandteilen einer Ontologie modelliert werden, ergeben sich verschiedene Definitionen von Ontologien im Forschungsgebiet formale Wissensrepräsentation. Wir führen hier die gängigsten Definitionen auf.
Eine weit verbreitete, frühe, jedoch für sich genommen wenig konkrete Definition von Ontologie liefert Gruber. Gruber [38] geht in seiner Definition einer Ontologie von einer abstrakten und vereinfachten Darstellung der Welt aus, einer Konzeptualisierung.

Definition (Ontologie nach Gruber): *Eine Ontologie ist eine explizite Spezifikation einer Konzeptualisierung.*

Der ungewöhnliche Terminus 'Konzeptualisierung' kann auch mit 'Begriffsbildung' übersetzt werden. Damit besagt die Definition, dass in einer Ontologie die Dinge der Welt oder eines Ausschnitts der Welt explizit aufgeführt werden. Das Wissen eines zu betrachtenden Bereiches wird in Grubers Anwendungen, die stark auf dem Knowledge Interchange Format (KIF) basieren, in einem deklarativen Formalismus dargestellt. KIF dient nach [55] als Sprache einer vermittelnden Schicht zwischen verschiedenen Implementierungen auf verschiedenen Plattformen. Diese Schicht formuliert das Wissen und den Kontext der Entwickler, welche nicht in ihren Implementierungen selbst festgehalten wurden, aber für eine Zusammenführung verschiedener Entwicklungen nötig sind. Die mit dem KIF repräsentierten Objekte des Wissensbereichs werden in einem Wörterverzeichnis wiedergege-

ben, auf das ein Programm, das Wissen darstellt und verarbeitet, zugreifen kann. Wir sehen an dieser Stelle, dass die konkrete Operationalisierung dieser Definition stark mit einer bestimmten formalen Sprache (KIF, siehe [55]) verbunden ist. Dies ist bei der folgenden Definition [93] nicht der Fall.

Definition (Ontologie nach Sowa): *Eine Ontologie ist das Ergebnis einer Studie der Kategorien von Dingen.*

Eine Ontologie ist in den weiteren Ausführungen Sowas ein Katalog der Typen von Dingen, die in einem bestimmten Bereich beziehungsweise Wissensgebiet existieren. Die Typen repräsentieren Prädikate, Wortsinn oder Begriffe und Beziehungstypen der Sprache, in der über die Dinge eines bestimmten Bereiches gesprochen wird. Die Kategorien können nach Sowa in einem Semantischen Netz angeordnet dargestellt werden und zwar so, dass im Netz unten liegende Kategorien durch die Kombination weiter oben liegender ersetzt werden können.
Während Sowas Definition und seine weiteren Ausführungen dazu die Frage nach den exakten Operationen, welche aus abstrakteren Begriffen konkretere Begriffe herstellen, aufwirft, kann die folgende Definition Guarinos als eine Übertragung der linguistischen Sicht Wittmers auf formale Sprachen statt natürliche Sprachen angesehen werden. In 'Understanding , Building, And Using Ontologies' [Gua97] erklärt Guarino Ontologien wie folgt:

Definition (Ontologie nach Guarino): *Eine Ontologie ist eine explizite, partielle Darstellung der intendierten Modelle einer logischen Sprache.*

Modelle der realen Welt werden in diesem Sinne mit Hilfe logischer Sprachen beschrieben. Diese Modelle werden durch Relationen zwischen den Begriffen aus der Welt aufgestellt. Für die Bedeutung der Relationen innerhalb einer Ontologie nach Guarinos Definition muss nun gelten, dass sie dieselbe bleibt, auch wenn die an der Relation beteiligten Instanzen ausgetauscht werden. Hierzu ein Beispiel: sitzen zwei Angestellte einer Bank, Herr Tischbein und Herr Stuhlbein, im selben Büroraum, und zwar so, dass Herr Tischbein am Fenster und Herr Stuhlbein am Gang sitzt, dann gilt dies auch aufgrund der allgemeinen Tatsache, dass ein Bankangestellter einen physikalischen Arbeitsplatz besitzt. Werden nun die Sitzplätze des Herrn Tischbein und des Herrn Stuhlbein vertauscht, dann besteht immer noch dieselbe allgemeine Tatsache, nämlich dass ein Bankangestellter einen physikalischen Arbeitsplatz besitzt. Die Relation zwischen 'Bankangestellter' und 'physikalischer Arbeitsplatz' wäre nach Guarinos Definition dann Teil der Ontologie.

Gangemi et al. [31] übernehmen die Definition der Ontologie von Guarino und fügen hinzu, dass auch Wörterbücher und Glossare Ontologien sind, allerdings in einem weniger formalen Grade. Ontologien können demnach nach ihrem Formalisierungsgrad unterschieden werden.

2.2.4 Ontologiedefinition der vorliegenden Arbeit

Die bis hier erläuterten Ontologiedefinitionen aus der Informatik weisen mehrere Nachteile auf, die durch die Ontologiedefinition für die vorliegende Arbeit aufzuheben sind.
Eine Sprachunabhängigkeit, das heißt, die Unabhängigkeit von einer bestimmten formalen Sprache, ist durch Grubers Arbeiten nicht gegeben. Die Sprachunabhängigkeit müsste durch eine Abstraktion des dort verwendeten KIF erst erarbeitet werden. Dies entspräche dann auch Grubers eigener Forderung, Ontologien mit einer minimalen Verzerrung durch ihre Kodierung zu definieren. Zudem wird bei Gruber nicht explizit auf die Ordnung der Begriffe eines Wissensgebietes eingegangen. Dies gilt auch für die Definition von Gangemi et al.
Guarino trifft in den Ausführungen zu seiner Definition eine Unterscheidung zwischen Begriffen und Instanzen des Begriffs. Eine Ober-Unterbegriffsrelation kann bei Guarino implizit aufgrund der beteiligten Instanzen und somit über die Begriffsextension gefunden werden. Nach den Untersuchungen von [73] ist dagegen die Unterscheidung zwischen Begriffen und Instanzen allerdings für Aussagen, die nicht in einer formalen Sprache getroffen werden, schwierig, weshalb wir in 1.2.1 gefordert hatten, Begriffe als Einheiten der Wissensrepräsentation zu betrachten. Sowa schließlich berücksichtigt zwar Begriffe als Einheiten und spricht auch von Ober-Unterbegriffsrelationen, seine Konkretisierungen beziehen sich allerdings auf eine allgemeine Ontologie logischer Typen nach Peirce[4].
Hieraus lassen sich abstrakte Anforderungen an eine Ontologie formulieren. Das Ziel ist dabei weiterhin die Einführung einer Ontologiedefinition, die von Menschen (Laien auf dem Gebiet der Wissensrepräsentation) und Maschinen lesbare und verarbeitbare Ontologien erklärt.

- Im Gegensatz zu Grubers Definition wird eine klare Identifikation der Bestandteile einer Ontologie gefordert.

- Anders als bei Guarino, Gruber und Gangemi ist eine eindeutige Identifikation der Bezüge dieser Bestandteile erforderlich.

- Insbesondere darf im Gegensatz zu Grubers und Guarinos Arbeiten

keine Bindung der Ontologiedefinition an eine formale Sprache erfolgen, wenn die Wissensrepräsentation praktisch handhabbar bleiben soll.

- Abweichend von Sowas Verständnis müssen fachgebietsspezifische Ontologien durch die Ontologiedefinition explizit miterfasst werden.
- Die Ordnung der Begriffe aus Abschnitt 2.2.1 muss Eingang in die Ontologiedefinition finden.

Eine Definition, die die Schwierigkeiten der vorhergehenden Definitionen umgeht und gleichzeitig Begriffe als Einheiten mit ihrer Ordnung und mit weiteren semantischen Relationen wie bei den im vorherigen Abschnitt vorgestellten semantischen Netzen behandelt, ist die Definition nach Stumme und Mädche. In ihrer Kurzform lautet sie [97]:

Definition 4 (Ontologie nach Stumme/Mädche) *Eine (Kern-) Ontologie ist ein 4-Tupel $\Omega := (B, \leq, R, \sigma)$, wobei B und R Mengen sind, $\leq$ eine transitive, reflexive und asymmetrische zweistellige Relation zwischen Elementen aus B, $\sigma : R \mapsto B \times B$ eine Relation, die jedem $r \in R$ Paare der Art $(B_1(r), B_2(r))$ mit $B_1 \subseteq B$ und $B_2 \subseteq B$ zuordnet. Wir nennen alle $b \in B$ Begriffe, $\leq$ die Unterbegriffsrelation, die Umkehrrelation der Unterbegriffsrelation notieren wir als $\geq$ und nennen sie Oberbegriffsrelation.*
Alle $r \in R$ nennen wir semantische Relationsnamen und $B_1(r)$ und $B_2(r)$ den Definitions- beziehungsweise den Wertebereich von r.

Die Forderung nach einer Ober-Unterbegriffsrelation, nach der Unabhängigkeit von einer formalen Sprache und nach Begriffen als Einheiten bei Beibehaltung der Vorzüge semantischer Relationen in semantischen Netzen ist mit dieser Definition gegeben. Die Definition unterscheidet die Menge der Relationsnamen von den eigentlichen Relationen. Dies weist bereits auf einen Aspekt der Ontologieerstellung hin: es soll bei einer möglichen Erweiterung einer bestehenden Ontologie Klarheit über die verwendbaren Relationen herrschen. Als Zusatz zur Kurzdefinition von Stumme und Mädche fordern wir im Sinne einer Verständlichkeit der Ontologie für den Menschen, dass jeder Begriff einen natürlichsprachlichen Namen besitzt und dass es einen (möglicherweise abstrakten) Wurzelbegriff gibt

Annahme 1 (Existenz des Wurzelbegriffs)

$$\exists \top \in B \forall b \in B : \; b \leq \top \tag{2.6}$$

Insgesamt stellt die Definition zusammen mit den Zusatzforderungen eine leichte Verkürzung der ausführlichen Ontologiedefinition der Karlsruher Gruppe in [97] dar.
Wir übernehmen diese Definition als Ontologiedefinition für den Rest der Arbeit. Um sie für angewandte Wissensrepräsentationen nutzbar zu machen treffen wir die Zusatzannahme, dass jeder Begriff mindestens ein Wort oder eine Wortfolge als natürlichsprachlichen Bezeichner besitzt. Diese Annahme liegt darin begründet, dass die Ersteller der Wissensrepräsentation durch Ontologien somit eine Möglichkeit erhalten, ohne formale Begriffsbildung, sondern durch die Umsetzung von Fachbegriffen zu Elementen der Menge B, einen wesentlichen Grundbestandteil der Ontologie zu erstellen. Es gilt für den Fortgang der Arbeit die

Annahme 2 (Existenz der Bezeichner) *Sei $\Omega := (B, \leq, R, \sigma)$ eine Ontologie. Dann gibt es für jedes $b \in B$ einen natürlichsprachlichen Bezeichner $D(b)$ in Form eines Wortes oder einer Wortabfolge natürlicher Sprache. Die Bezeichner für die Begriffe einer Ontologie Ω notieren wir auch als $D(\Omega)$.*

In den Anwendungen der vorliegenden Arbeit weisen Bezeichner eine Länge von wenigen Wörtern auf, die Verwendung von Textabschnitten oder Texten als Begriffsbezeichner schließen wir aus pragmatischen Gründen aus. Die grafische Repräsentation der Ontologien nach dieser Definition findet sich in der Beispieldarstellung 2.4.

Jede gerichtete Linie ohne Beschriftung kann als Unterbegriffsrelation gelesen werde, jedes kursiv gedruckte Wort als Begriff, jede beschriftete Linie als Relation zwischen zwei Begriffen. Zusätzliche gerichtete Linien, deren entsprechende Relationen etwa durch die Halbordnungseigenschaft der Unterbegriffsrelation zwischen $\top$, dem abstrakten Wurzelbegriff, und *Lehmann* gelten, werden der Übersicht halber nicht in das Diagramm eingetragen.
Die Definition 4 erfüllt zusammen mit den Anforderungen 1 und 2 die am Anfang des vorliegenden Abschnitts aufgestellten Bedingungen. Alle Bestandteile einer Ontologie und ihr Verhältnis zueinander sind erklärt, die Definition ist nicht an einen speziellen Wissensrepräsentationsmechanismus oder eine bestimmte formale Sprache gebunden und die Reflexivität, Asymmetrie und Transitivität der Unterbegriffsrelation $\leq$ ist per se in der Definition enthalten. Die Einführung der natürlichsprachlichen Bezeichner gewährleistet die fachgebietsspezifische Anwendung auch für Nichtinformatiker. Unter einer **fachgebietsspezifischen Ontologie** verstehen wir für den Rest der vorliegenden Arbeit eine Ontologie im Sinne der Definition 4 und der Zusatzannahmen, die Fachwissen repräsentiert. Eine fachgebietsspezifische

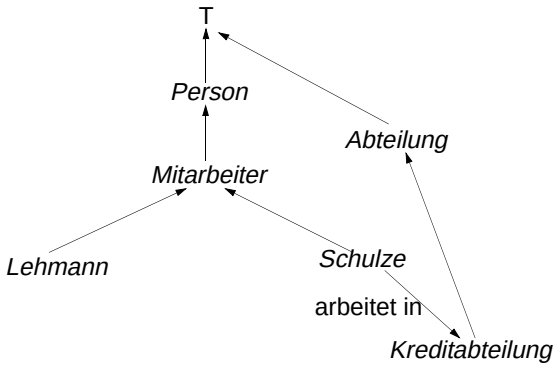

Abbildung 2.4: Grafische Darstellung einer Ontologie

Ontologie besitzt somit einen begrenzten Gültigkeitsbereich und wird von Experten des betreffenden Fachgebietes erstellt.

2.3 Ontologieanwendungen

Im folgenden Abschnitt stellen wir Anwendungsgebiete der Wissensrepräsentation durch Ontologien vor. Diese Anwendungen weisen das Charakteristikum auf, dass sie auf Begriffe aus Ontologien als Vokabular zurückgreifen. Ontologieanreicherungsverfahren werden im Laufe dieser Arbeit dazu dienen, dieses Vokabular zu erweitern und als Begriffsvorschlag innerhalb einer bestehenden Ontologie semantisch einzuordnen.
Es existieren mehrere Klassifikationsansätze für die Anwendungsfelder von Ontologien. In diesem einleitenden Abschnitt werden wir Teile der Klassifikationsansätze übernehmen, wobei wir uns danach richten, welche Anwendungen nach der für diese Arbeit gefundenen Ontologiedefinition denkbar oder bereits vorhanden sind.
McGuinness untersucht in [69] die Anwendungsfelder von Ontologien ausschließlich aus der Perspektive, die sich mit der wachsenden Forschergemein-

schaft rund um das Themengebiet 'Semantic Web' herausbildet. Die Unterscheidung, die sie dabei trifft, richtet sich nach dem Formalisierungsgrad der untersuchten Ontologien. Diese werden dabei nach leichtgewichtigen und schwergewichtigen Ontologien unterschieden, wobei die Grenze zwischen diesen anhand der Formalität der Unterbegriffsrelation zu ziehen ist. Wenn es ein formales Kriterium für die Bildung einer Unterbegriffsrelation zwischen zwei Begriffen in einer Ontologie gibt, dann kann in diesem Sinne von einer schwergewichtigen Ontologie gesprochen werden. Ein Beispiel für ein solches Kriterium wurde mit Definition 3 gezeigt. Formale Begriffe zusammen mit der formalen Unterbegriffsrelation aus Abschnitt 2.2.1 ergeben dann eine schwergewichtige Ontologie und Ontologie nach Definition 4, wenn zu jedem formalen Begriff auch noch ein natürlichsprachlicher Begriffsbezeichner vorhanden ist[1].

Bei leichtgewichtigen Ontologien muss die Unterbegriffsrelation keine weitere äußere Begründung erfahren, und genau das ist bei Definition 4 der Fall. In den Beispielen der Arbeit werden wir somit aus McGuinness' Klassifikation nur diejenigen Anwendungsfälle übernehmen, die eine informelle Unterbegriffsbeziehung zulassen.

Besagte Anwendungsszenarien sind auch auf in McGuinness Sinne formale Ontologien ausdehnbar. Die aus [69] verbleibenden Anwendungsgebiete sind somit Navigation, Suche und Sinnunterscheidung aus der Sicht der Benutzer, die keine Ontologie erstellen oder eine ontologieunterstützte Anwendung implementieren. Diese Anwendungsgebiete sind grundsätzlich auch für Datenbestände denkbar, die lokal oder innerhalb einer begrenzten Benutzergruppe zugänglich sind.

Eine weitere Klassifikation von Ontologieanwendungen stammt von Uschold et al. [101] und wurde von van Zyl et al. [102] erweitert und mit Beispielen versehen. Sowohl die Definition von Ontologie als auch die beschriebenen Anwendungen sind sehr allgemein erläutert, was sich auch darin äußert, dass van Zyl hauptsächlich die Architekturen der klassifizierten Systeme zeigt, während ontologiebasierte Abläufe innerhalb der Systeme nicht im Detail ausgeführt werden. Zu den von bisher aufgeführten Anwendungsgebieten kommen nach Uschold und van Zyl noch die Nutzung von Ontologien als Spezifikation, was als Erweiterung der Bestimmung eines fachspezifischen Vokabulars bei McGuinness angesehen werden kann. Ein wichtiges zusätzliches Klassifikationskriterium, das van Zyl auch über die von Uschold bestimmten Kategorien hinaus identifiziert, ist der Einsatz mehrerer Ontologien in Suchanwendungen. Sowohl dieses zusätzliche Klassifikationskriterium

[1] $\top$ ist dann vorhanden und R möglicherweise leer.

als auch die Nutzung von Ontologien bei der Spezifikation sind mit Definition 4 und den Zusatzforderungen vereinbar.
Bei Ontologien, die eine formalere Unterbegriffsdefinition aufweisen, ergeben sich bei allen bestehenden Klassifikationsansätzen auch noch Anwendungen, bei denen die Auswertung einer Ontologie Fragen zu einem Wissensgebiet beantwortet, indem bestehende Relationen genutzt werden. Diese Anwendungen können unabhängig von einer Suche nach Ressourcen (beispielweise Dokumenten) in einem Informationsbestand verlaufen. Solche Anwendungen werden in unseren Beispielen zur Suche Berücksichtigung finden, da auch für Ontologien nach Definition 4 bei einer entsprechend klaren Formulierung der Relationen aus R sinnvolle Antworten geschlossen werden können.
Die für die vorliegende Arbeit relevanten Ontologieanwendungen verbinden stets einen Datenbestand, der in Form von elektronischen Dokumenten vorliegt, mit dem in der Ontologie wiedergegebenen Wissen. Wir halten aus der Betrachtung der bestehenden Klassifikationen von Ontologieanwendungen fest, dass für unsere Beispiele in diesem Abschnitt folgende Anwendungsgebiete in Frage kommen:

- Suche in (möglicherweise heterogenen) Datenbeständen mit Hilfe einer oder mehrerer Ontologien
- Navigation in (möglicherweise heterogenen) Datenbeständen mit Hilfe einer oder mehrerer Ontologien
- Spezifikationen, bei denen Bestimmungen des Wortsinns durch Ontologien unterstützt werden

Suche und Navigation unter Zuhilfenahme einer Ontologie weisen eine enge Verwandtschaft auf, und wir werden sie daher in einem gemeinsamen Abschnitt behandeln.
Die weiteren Anwendungsgebiete nach der obigen Einteilung stellen wir im Anhang dar.

2.3.1 Suche und Navigation

Die ontologiebasierte Suche nutzt die Eindeutigkeit der in der Ontologie festgehaltenen Beziehungen. Eine typische Suche in diesem Sinne schließt von dem Nutzer bekannten Begriffen über Verallgemeinerungen und Spezialisierungen (das heißt über die Ober-Unterbegriffsrelation) sowie über die anderen Relationen auf unbekannte Begriffe. Der prinzipielle Ablauf dieser Suche unterscheidet sich für begrenzte Datenbestände oder verteilte und offene Archivierung nicht prinzipiell. Wir beziehen die Beispiele in diesem

Abschnitt auf zwei unterschiedliche Szenarien, nämlich die Pflege eines Firmenwissens und die fortgeschrittene Anwendung herkömmlicher Internet-Suchmaschinen wie Google. Der Datenbestand einer Firma wird in den Ausführungen als abgegrenzt von anderen betrachtet, während eine Suchmaschine wie Google prinzipiell auf offene, sich kontinuierlich durch die Aktivität nicht explizit vordefinierter Teilnehmer erweiternde Datenbestände zugreift.
Wissen wird in wirtschaftlichen Abläufen zunehmend als eigener Ressourcentyp neben den klassischen Ressourcen (nach [21] Arbeitskraft, Boden, Kapital) bewertet. Dementsprechend entstehen Gegenentwürfe (vergleiche [8] und [54]) zur klassischen Theorie des Unternehmens [15] mit hierarchischen Organisationsstrukturen und einem in den oberen Stufen der Firmenhierarchie konzentrierten Wissen um die Produktion von Gütern und Dienstleistungen. Vielmehr muss nach diesen Gegenentwürfen ein Unternehmen systematisch die interne Weitergabe firmeninternen Wissens pflegen.
Die erstrebte Kommunikation des Firmenwissens kann beispielsweise durch eine Ontologie erleichtert werden. Die Ontologie kann das nötige Vokabular für firmeninterne Dokumentationen liefern. Das Vokabular kann entweder im Dokumentationstext selbst oder als Schlagwort, das einem Teil des Dokumentationstextes zugeordnet wird, verwendet werden. Durch das Betrachten der Ontologie oder durch eine an die Ontologie gerichtete Anfrage können dann die für das Firmenwissen charakteristischen Suchbegriffe (im Sinne der Definition die Menge B) und ihre inhaltlichen Verknüpfungen (im Sinne der Definition 4 die Ausprägung der Menge R durch σ) identifiziert werden. Durch die firmeninterne Standardisierung, zu der die Ontologie in diesem Falle beiträgt, wird die Suche erleichtert, da eine größere Transparenz über Terminologien und Zusammenhänge hergestellt werden kann. Insbesondere bietet dies Vorteile bei der Einarbeitung, bei der Überschreitung von Fachgrenzen und bei der Schulung von Mitarbeitern.
Betrachten wir im Vergleich dazu Suchprozesse, die aus der Nutzerinteraktion mit einer Anwendung wie Google entstehen. Suchmaschinen sind, da nur ein geringer Teil der Webdokumente in Kategorien wie bei Google oder Yahoo eingeordnet wird, auf Freitexteingaben von Suchstrings angewiesen. Problematisch wird eine Sucheingabe dann, wenn der genaue Suchterm nicht bekannt ist. Beispielsweise könnte für einen angehenden Wirtschaftsjournalisten bei der Recherche nach einem Artikel interessant sein, welche Firmenbeteiligungen in einem bestimmten Wirtschaftszweig existieren, um danach mit einer herkömmlichen Suchmaschine aktuelle Artikel über die wirtschaftlich miteinander verflochtenen Firmen zu finden.
Für beide Fälle lässt sich die selbe Abstraktion eines ontologieunterstützten

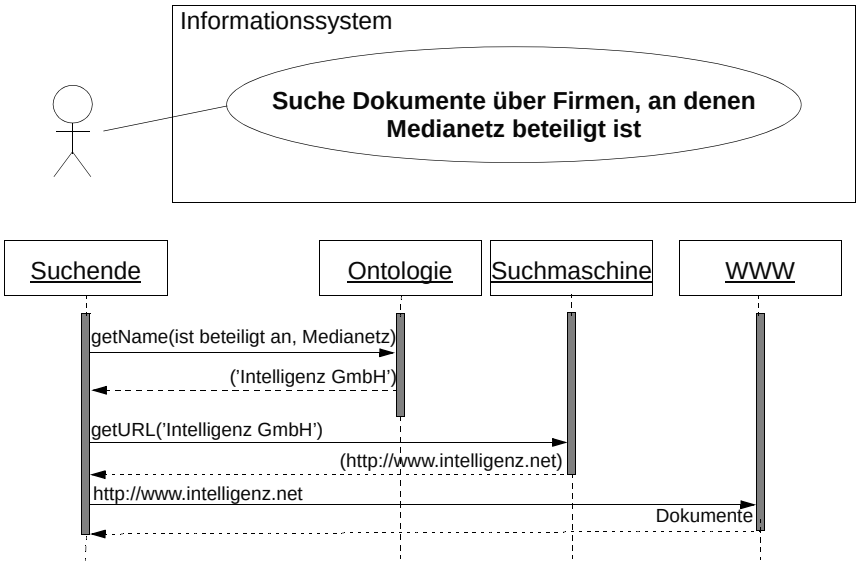

Abbildung 2.5: Ontologien als Unterstützung bei Suchen

Suchprozesses angeben, wie sie anhand des Sequenzdiagramms in Abbildung 2.5 deutlich wird. Hier ist es das Ziel der Suche, im WWW verfügbare Dokumente über Firmen, an denen die Firma Medianetz beteiligt ist, zu finden. Der Nutzer richtet zunächst eine Suchanfrage an die Ontologie, die begriffliche Zusammenhänge und eine Menge von Begriffsnamen zurückliefert. Die Suchanfrage an die Ontologie wird hier durch die zweistellige Methode getName(Relation, Begriff) versinnbildlicht. Für die Antwort ('Intelligenz GmbH') gilt bei gegebener Ontologie $\Omega := (B, \leq, R, \sigma)$ und 'ist beteiligt an' $\in R$:

$$\{Intelligenz\ GmbH\} = \{b \in \Omega | \text{'ist beteiligt an'}(Medianetz, b) \in \sigma\} \quad (2.7)$$

Mit dem Suchstring ('Intelligenz GmbH') ist dann eine Suchanfrage an Google möglich, die die relevanten Adressen im WWW zurückliefert. Bei

einer Suche ohne Ontologie wäre entweder ein Vorwissen beim Nutzer oder eine Vorrecherche durch Erfassung der Google-Suchergebnisse zu Medianetz nötig gewesen. Die ontologieunterstützte Suche läuft ähnlich ab, wenn nicht Dokumenteninhalte, sondern den Dokumenten zugeordnete Schlagwörter eines firmeninternen Wissensmanagementsystems ausgewertet werden. In diesem Falle richtet sich die zweite Anfrage nicht an eine Suchmaschine, sondern an eine interne Datenbank, in der die Verschlagwortungen festgehalten sind. Der Übergang von Suchvorgängen zu Navigationsvorgängen ist fließend. Wäre im obigen Beispiel auch Medianetz Ergebnis einer Suche, die graphisch unterstützt ablief, so könnte dies schon als Navigationsanwendung betrachtet werden. Wir halten fest, dass alle Suchvorgänge, die eine graphische Repräsentation der Ontologie benutzen, auch Navigationsvorgänge sind. Ebenso fließend sind die Übergänge zwischen Suche, Navigation und dem Beantworten fachlicher Fragen durch die Ontologie. Der prinzipielle Unterschied besteht lediglich darin, dass die Ontologie an sich bei der reinen Beantwortung fachlicher Fragen zu Rate gezogen wird, ohne das externe Dokumente auf ihren Volltext oder ihre Verschlagwortung hin überprüft werden. Bei einer solchen Anwendung würde das Sequenzdiagramm 2.5 bereits nach der Abfrage der Ontologie enden. Auch hier kann die Abfrage aus mehreren aufeinander folgenden Fragen bestehen.
Der Einsatz mehrerer Ontologien bei einer in der Teilnehmerzahl beschränkten Suchanwendung kann anhand der Abbildung von Produktkatalogen gezeigt werden, ontologiebasierte Spezifikationen anhand der patternbasierten Sicherheit in der Informations- und Kommunikationstechnik. Als Vertreter von Anwendungen mit einer per se unbeschränkten Teilnehmerzahl werden wir im Anhang den Ansatz des 'Semantic Web' und des 'Web of People' besprechen. Bei beiden verschränken sich die genannten Anwendungsfelder. Für diese weiteren mit Suche und Navigation verwandten Ontologieanwendungen verweisen wir auf den Anhang.

2.4 Zusammenfassung

In diesem Kapitel wurden Ontologien und ihre Anwendungen beschrieben. Wir haben dabei eine Ontologiedefinition herausgearbeitet, die Wissensrepräsentation nahe an die natürliche Sprache rückt und dadurch auch für Nichtinformatiker handhabbar werden lässt. Dazu wurden eine Unterbegriffsrelation, die eine Halbordnungsrelation darstellt, sowie die Existenz eines natürlichsprachlichen Begriffsnamens für jeden Begriff in der Ontologie und die Existenz eines abstrakten Wurzelbegriffes in die Definition

aufgenommen. Diese Strukturelemente sind nicht an eine bestimmte formale Sprache, in der die Ontologie verfasst ist, gebunden.
Zudem haben wir verschiedene Anwendungsgebiete von Ontologien aufgeführt, die mit der vorausgegangenen Ontologiedefinition entwickelt werden können. Die ontologiebasierte Suchunterstützung erscheint dabei die umfassendste Art der dargestellten Anwendungen zu sein, da sie die Grundlage für die anderen Anwendungstypen liefert. Wir haben zudem erläutert, dass Ontologien sowohl in lokalen als auch in globalen Anwendungen der Wissensrepräsentation eingesetzt werden können.
Im Folgenden werden wir ausschließlich fachgebietsspezifische Ontologien betrachten, da für diese besondere Probleme bei ihrer Erstellung auftreten.

Kapitel 3

Ontologieerstellung

Im folgenden Kapitel werden wir uns mit dem Problem der Ontologieerstellung und ihrer möglichen automatischen Unterstützung befassen.
Zunächst werden wir definierte technische Abläufe untersuchen, die der Erstellung fachgebietsspezifischer Ontologien dienen. Diese Prozesse der Ontologieerstellung gehen von der Annahme aus, dass Wissensrepräsentation ähnlich wie Spezifikationen [75], oder die Softwareentwicklung [34] von einer genauen Festlegung der Arbeitsschritte, der Arbeitsteilung und der Quellen der Repräsentation stark profitiert. Gruppenpsychologische oder soziale Aspekte der notwendigen Interaktion werden hier nicht betrachtet.
Nach der Untersuchung verschiedener Ontologieerstellungsprozesse werden wir den Einsatz eines bestimmten Prozesses in der Praxis des Forschungsprojektes k-med [43] vorstellen. Die Anwendung und detaillierte Ausformulierung des Ontologieerstellungsverfahrens in k-med wurde durch den Verfasser der vorliegenden Arbeit durchgeführt.
Wir schließen die Untersuchungen des vorliegenden Kapitels mit einer Anforderungsanalyse ab. Diese widmet sich der Frage, wie die vorgestellten Verfahren und insbesondere das in k-med eingesetzte Verfahren durch eine automatische Unterstützung des Ontologieerstellungsprozesses verbessert werden können. Wir entwickeln Anforderungen an die automatische Unterstützung von Ontologieerstellungsprozessen durch Methoden, die für Ontologien nach Definition 4 und den Zusatzannahmen 1 und 2 geeignet sind. Des Weiteren spezifiziert der Schlussteil des folgenden Kapitels Anforderungen an automatische Verfahren, die bereits bestehende Ontologien erweitern.
Der erste Abschnitt des vorliegenden Kapitels erklärt die Notwendigkeit definierter Ontologieerstellungsprozesse und liefert eine Kategorisierung gängiger Verfahren. Die Abschnitte 3.2 bis 3.5 stellen gängige Methoden vor.

Die Ontologieerstellung im k-med Projekt wird in Abschnitt 3.6 erläutert. Anhand der detaillierten Ausformulierung und der Erfahrungen aus k-med erfolgt danach die Anforderungsanalyse zur automatischen Unterstützung von Ontologieerstellungsprozessen in Abschnitt 3.7.

3.1 Ontologieerstellungsprozesse

In diesem Abschnitt gehen wir auf Ontologieerstellungsprozesse und ihre grundlegenden Ansatzpunkte ein.
Ein wesentliches Problem für den Einsatz von fachgebietsspezifischen Ontologien liegt darin, dass für bestimmte Projekte und Einsatzszenarien die jeweilige Ontologie nicht vorhanden ist oder nicht zur Verfügung steht. Letzteres kann beispielsweise dann der Fall sein, wenn zwar eine Ontologie existiert, die dem Projekt dienlich wäre, diese aber nur gegen Bezahlung zu nutzen ist. Das Cyc-Projekt etwa [61] stellt nur die Begriffe auf den abstrakteren Ebenen der zugehörigen Cyc-Ontologie $\Omega_{cyc} := (B_{cyc}, \leq, R, \sigma)$ zur Verfügung. Nur eine Menge B_{sub} von Begriffen, für die für alle $b \in B_sub$

$$|\{b_0 \in B_{cyc} | b_0 \geq b\}| \tag{3.1}$$

klein ist, kann kostenlos für Anwendungsentwicklungen genutzt werden.
Die Erstellung einer fachgebietsspezifischen Ontologie durch eine einzelne Person gewährleistet nicht, dass auch tatsächlich ein fachlicher Konsens in die Modellierung eingeht. Wie die Beispiele aus den Abschnitten zur Anwendung der Wissensrepräsentation durch Ontologien aus dem vorangegangenen Kapitel dieser Arbeit zeigen, wird ein solcher Konsens aber stets notwendig sein, um die Anwendung einer Ontologie über einen individuellen Bereich hinaus gewährleisten zu können.
Die manuelle Erstellung fachgebietsspezifischer Ontologien gestaltet sich somit zeit- und kostenintensiv, da eine Anzahl von Experten des zu modellierenden Fachgebietes für diese Aufgabe intensiv zusammenarbeiten muss. Ferner sind nicht alle Beteiligten mit den Formalismen der Ontologieerstellung vertraut. Insgesamt ergibt sich ein hoher Kommunikationsaufwand, da den formal geschulten Personen, die einer Expertengruppe Ontologieerstellung nahe bringen können (in der Regel handelt es sich hierbei um Informatiker, Mathematiker oder Personen aus einem verwandten Gebiet), das Fachgebiet der Ontologie fremd ist. Hinzu tritt ein weiteres prinzipielles Problem: bei zunehmender Größe einer Ontologie lässt die Überschaubarkeit der Gesamterstellung nach.
Dieser Widerspruch lässt sich nur auflösen, wenn bei der Ontologieerstel-

lung ein Ablaufschema eingehalten wird. Solche Erstellungsprozesse sind bei der Softwareentwicklung erprobt und erfolgreich [75], [34]. Der Hauptunterschied zu Ontologieerstellungsprozessen liegt, selbst im Vergleich zu einer nutzerzentrierten Anforderungsanalyse bei der Softwareentwicklung, in den inhaltlichen Rollen der Fachgebietsexperten. Bei der Entwicklung einer fachgebietsspezifischen Ontologie ist eine noch direktere Repräsentation des Expertenwissens von zentraler Bedeutung. Wir werden im Folgenden verschiedene Verfahren der systematischen Ontologieerstellung in Abgrenzung zu intuitiven Verfahren vorstellen und anschließend bewerten. Gemäß unserer Ontologiedefinition (Definition 4 aus Kapitel 2) haben wir Verfahren ausgewählt, die unabhängig von einer konkreten Darstellungssprache wirken. Hauptbewertungskriterien für den praktischen Einsatz der Verfahren werden der Nutzeraufwand, der mit einem Verfahren verbunden ist und die Erweiterbarkeit des Verfahrens um eine automatische Unterstützung der Ontologieerstellung sein. Wir verfolgen damit das Gesamtziel einer Erleichterung und Beschleunigung des Ontologieerstellungsprozesses. Die intensive Zusammenarbeit der Fachgebietsexperten soll dabei strukturiert verlaufen, der Prozess soll sich an der späteren Anwendung der Ontologien orientieren und der Schulungsaufwand zum Vermitteln des Prozessablaufs soll geringer sein als der Nutzen, der sich ergibt.
Nach [45] können können Verfahren zur Ontologieerstellung als intuitive[1], induktive, deduktive, synthetische und kooperative Verfahren kategorisiert werden.
Bei intuitiven Verfahren geht der Entwickler der Ontologie von der Anwendung der Ontologie aus. Er lässt subjektive Faktoren (wie etwa persönliche Kreativität und eigene Vorstellungen bezüglich des Interessengebietes) in die Ontologieentwicklung einfließen. Sofern der Entwickler hierbei nicht gleichzeitig die Fähigkeit besitzt, Vorschriften oder eine Systematik zur Ontologieentwicklung vorzugeben, um seine Einsichten auf eine Gruppe von an der Erstellung beteiligten Fachexperten zu übertragen, kann das Resultat einer intuitiven Entwicklung einer starken Begrenzung unterliegen. Gute Resultate lassen sich hier nur dann erzielen, wenn die persönlichen Ansichten des Entwicklers von vornherein im Konsens mit der Ansicht einer größeren Gruppe von Experten stehen.
Wir können aufgrund dieser Betrachtungen intuitive Verfahren als wenig praxistauglich für einen Gesamtprozess der Ontologieerstellung ansehen und

[1]Die wörtliche Übersetzung würde 'inspirativ' lauten. Da sich die Autoren auf eine spontane Inspiration, die in die Ontologieerstellung eingeht, wurde für vorliegende Arbeit die im Deutschen gängigere Wortwahl 'intuitiv' verwendet.

werden auch keine ausschließlich intuitiven Verfahren untersuchen. Allerdings bleibt festzuhalten, dass viele Ontologieerstellungsverfahren in ihren ersten Phasen Techniken des Brainstormings [79] einfließen lassen oder zumindest zulassen. In diesem Rahmen werden intuitive Elemente folglich als eine zu strukturierende Grundlage beibehalten.
Es verbleiben nach [45] insgesamt vier weitere Kategorien von Erstellungsprozessen, die im Folgenden ausgeführt werden.

3.1.1 Kategorien von Erstellungsprozessen

Eine Kategorisierung der Ontologieerstellungsprozesse erleichtert die Beurteilung ihrer Praxisrelevanz für den im Verlaufe des Kapitels darzustellenden Anwendungsfall.
Die verbleibende Unterteilung von Ontologieerstellungsmethoden nach Holsapple und Joshi unterscheidet vier Kategorien. Diese Kategorien können in der Praxis nicht strikt voneinander getrennt werden, vielmehr bilden sie bei den zu untersuchenden Erstellungsprozessen Kernaspekte des Vorgehens.
Bei so genannten **induktiven Erstellungsmethoden** (Induktion: das abstrahierende Schließen vom Speziellen auf das Allgemeine) wird eine Ontologie durch Beobachtung, Überprüfung und Analyse spezifischer, in der Realität vorliegender Fälle im durch die Ontologie zu erfassenden Wissensgebiet erstellt. Die daraus resultierende Ontologie, deren Gültigkeit für bestimmte Fälle sichergestellt wurde, muss danach durch ihre Anwendung auf andere Fälle des gleichen Wissensgebietes überprüft und gegebenenfalls verändert werden.
Parallelen sind bei der induktiven Ontologieerstellung insbesondere zum 'case based reasoning', dem fallbasierten Schließen [2], zu sehen. In der Vorgehensweise der im vorherigen Kapitel im Unterabschnitt 2.2.1 eingeführten formalen Begriffsanalyse würde das Grundprinzip hier darin bestehen, von einem formalen Kontext $K := (G, M, I)$ auf die Begriffe (die nicht identisch sein müssen mit der Gesamtheit der aus K ableitbaren formalen Begriffe) und ihre semantischen Relationen zu schließen. Die Gültigkeit des resultierenden Modells muss dann bei einer Erweiterung des formalen Kontexts durch ein $K' \supseteq K$ untersucht werden.
Im Gegensatz zur induktiven Methode steht die **deduktive Methode** (Deduktion: Schließen von Allgemeinen auf das Besondere). Hier werden zunächst allgemeine Regeln zur Konstruktion der Ontologie angegeben, die abstrakteren Begriffe der Ontologie angelegt und dann mit Hilfe des Regelwerks immer speziellere Unterbegriffe erstellt. Durch die Spezialisierung erfolgt dabei

eine sukzessive Annäherung an das eigentliche Wissensgebiet. Ein Beispiel für ein Regelwerk könnte darin bestehen, dass mit der Einführung eines Unterbegriffes anhand bestimmter Merkmale auch stets der Begriff eingeführt werden sollte, der diese Merkmale nicht erfüllt. Mit dem Begriff *pathogenes Bakterium* als Bakterium, das Krankheiten verursachen kann, sollte beispielsweise auch der Begriff *nicht pathogenes Bakterium* als Bakterium, das nie Krankheiten verursacht, eingeführt werden. Durch die Festlegung des deduktiven Regelwerks zur Ontologieerstellung folgt, dass für alle Begriffe stets auch eine (mehr oder minder) formale Definition vorliegt.
Bei der **synthetischen Erstellungsmethode** gewinnt man aus einer Menge vorhandener Ontologien $\{\Omega_1 := (B_1, \leq, R, \sigma),...,\Omega_n := (B_n, \leq, R_n, \sigma)\}$ eine Gesamtontologie. Für diese Ausgangsontologien sollten möglichst wenige semantische Überschneidungen vorliegen, das heißt idealerweise eine kleine Mächtigkeit

$$|\bigcap\nolimits_{1\leq i\leq n} B_i| \tag{3.2}$$

da ansonsten Probleme bei der Abbildung, Unterscheidung und Vereinigung ähnlicher Begriffe auftreten können, die die Erstellung der Gesamtontologie erschweren. Es handelt sich hierbei um eine Idealvorstellung. Aufgrund der unterschiedlichen Prämissen, unter denen die Teilontologien entwickelt wurden, geht der Akt der Zusammenführung weit über die Vereinigung der Begriffsmengen und die Einführung zusätzlicher Unterbegriffsrelationen auf einer abstrakten Ebene der Ontologie hinaus. Wenn beispielsweise eine der ursprünglichen Ontologien nach dem deduktiven Prinzip entwickelt wurde, so steht die Entscheidung an, ob die Deduktionsregeln auf die anderen bei der Synthese beteiligten Ontologien im Sinne einer einheitlichen Modellierung ausgedehnt werden sollten. Ein solcher wechselseitiger Einfluss der Erstellungsprinzipien kann sowohl positiv als auch negativ sein.
Mit der gemeinschaftlichen oder auch **kooperativen Ontologieentwicklung** stellen Holsapple und Joshi auf das Zusammenwirken der Entwickler bei der Ontologieerstellung ab. Unterschiedliche fachliche Erfahrung und persönliche Standpunkte fließen dabei in die Erstellung mit ein. Der Vorteil einer solchen Entwicklungsmethode liegt in einer erhöhten Akzeptanz der entstehenden Ontologie über die individuelle Sicht eines Beteiligten hinaus. Hier sind wiederum Parallelen zur Softwareentwicklung zu sehen: in der objektorientierten Modellierung wird die Aufstellung eines gemeinsamen Modells der bei der Implementierung (für eine Beispielanwendung siehe [24]) zu verwendenden Klassen und Instanzen priorisiert. Alle Entwickler sind an dieser Modellierung beteiligt. Im so genannten Extreme Programming-Ansatz [64] erfolgt sogar eine ständige produktive wechselseitige Kontrolle

der Entwickler, es ist gleichsam untersagt, alleine zu entwickeln. Der durch solche Verfahren etablierte Konsens ist allerdings nur über klar definierte Prozesse und klar definierte Verantwortlichkeiten zu erreichen. Es besteht keine Grenze zur induktiven, deduktiven oder synthetischen Methode.
Zusätzlich zu den in der Klassifikation von Holsapple und Joshi angegebenen Charakteristika haben wir in diesem Abschnitt Beispiele und eine algebraische Schreibweise gemäß der Definition 4 angewandt. Insbesondere halten wir fest, dass es bei der synthetischen und bei der gemeinschaftlichen Erstellung Wechselwirkungen mit den anderen Erstellungsprinzipien gibt. Diese sind bei Holsapple und Joshi in dieser Form nicht explizit aufgeführt. Die tatsächlichen Ontologieerstellungsmethoden stellen nun auf eine genauere Spezifikation der Quellen, der Aktivitäten und der einzelnen Beteiligten ab. Die oben beschriebene Kategorisierung kann hierbei eher als eine Abstraktion der Aktivitäten bei den Erstellungsprozessen angesehen werden. Wir werden diese Aktivitäten bei der Darstellung der Entwicklungsmethoden stets berücksichtigen. Es folgen die Methoden nach Uschold, Gangemi, Fernandez-Lopez/Gomez-Perez, und Holsapple/Joshi. Diese Methoden wurden aus der weit größeren Anzahl von Erstellungsmethoden (vergleiche beispielsweise [36]) ausgewählt, weil sie allesamt eine Erklärung über die Quellen der Begriffe und die Konstruktion einer initialen Ontologie enthalten.

3.2 Ontologieerstellung nach Uschold

Mike Uschold [101] unterscheidet zwei Hauptphasen der Ontologieerstellung, nämlich die informelle Phase und die Phase, in der eine Formalisierung der Ontologie vorgenommen wird. Es besteht zwar die Möglichkeit, ohne vorherige Planung durch eine informelle Phase direkt eine Formalisierung durchzuführen, Uschold rät jedoch davon ab. Eine weitere Verfeinerung beim Übergang von der informellen Phase zur Formalisierung bildet die Erstellung einer informellen Ontologie, auf deren Grundlage dann die Formalisierung vorgenommen werden kann. Am Ende der Ontologieerstellungsphasen steht immer eine Evaluation, entweder als Bewertung der Ergebnisse der abschließenden Formalisierung oder als Bewertung der möglicherweise vorliegenden informellen Ontologie.
Betrachten wir nun die informelle Vorbereitung. Innerhalb der informellen Phase kommt es zu einer Eingrenzung der Anwendung durch die Benutzergruppe, der die zu erstellende Ontologie dienen soll. Die Gruppe der Benutzer formuliert vom Allgemeinen zum Speziellen Anwendungsszenarien

in Form von so genannten Kompetenzfragen. Im Falle einer medizinischen Ontologie könnte dies beispielsweise die Frage beinhalten, ob die Ontologie vordringlich zu Lehrzwecken oder zur Diagnoseunterstützung dienen soll. Wenn eine Entscheidung für einen stärker auf die Didaktik bezogenen Einsatz fällt, so müsste eine weitere Spezifikation der Anwendungsfälle die Frage beantworten, welche Lernziele erreicht werden sollen und nach welcher didaktischen Methode der Unterrichtsstoff vermittelt werden soll. Eine weitere Kompetenzfrage beträfe in diesem Beispiel dann den Detaillierungsgrad der Ontologie: soll sie selbst dazu in der Lage sein, auf Fragen eines Lernenden Antworten zu liefern oder dient sie als Grundlage zur Verschlagwortung von Unterrichtsmaterialien? Bei einer anderen Grundsatzentscheidung, die die Ontologie als Diagnose- und Therapieunterstützung vorsieht, muss die Interaktion zwischen behandelndem Arzt und der ontologiebasierten Wissensrepräsentation spezifiziert werden. Die Verfeinerung der Kompetenzfragen kann als Gegenstück zu den Begriffen und ihre Ober-Unterbegriffsbezüge, die später in die eigentliche Ontologie eingehen, angesehen werden. Daher verläuft die Vorgehensweise an dieser Stelle deduktiv.
Das Vokabular für die Begriffe beziehungsweise ihre jeweiligen Bezeichner kann durch Brainstormingtechniken [79] gewonnen und den Kompetenzfragen angepasst werden. Das Brainstorming kann in diesem Zusammenhang als intuitiver Aspekt der von Uschold vorgestellten Methode angesehen werden. Die durch die Anpassung gewonnenen Begriffe bilden die Grundlage für die formale Phase. Erst in dieser werden Techniken einer formalen Ontologiesprache zu Rate gezogen. Die formale Umsetzung fällt in den Aufgabenbereich von Experten der Wissensrepräsentation, kooperative Elemente sind bei Uschols Methode vor allem während der informellen Phase zu erkennen. Insbesondere erfolgt zwischen den Anwendern der Ontologie und den Erstellern eine enge und schriftlich fixierte Absprache über den (in Form von Fragen formulierten) Kompetenzbereich jedes einzelnen Begriffs. Induktive und synthetische Aspekte werden von Uschold nicht ausdrücklich erwähnt.

3.3 Ontologieerstellung nach ONIONS

Gangemi et al. [31] beschreiben das Problem der Ontologieerstellung in erster Linie als Relevanz- und Eingrenzungsproblem, als so genanntes 'stopover problem', vergleiche für den medizinischen Anwendungsbereich [41]. Das 'stopover problem' entsteht dann, wenn keine eindeutige Regel besteht, wie genau und wie ausgedehnt die Ontologie modelliert werden soll. Das Ziel der Vorgehensweise Gangemis besteht in einer Ontologie, die keine überflüssigen

Teile beinhaltet und daher nur dort begriffliche Unterscheidungen trifft, wo es notwendig ist. Da es sich bei ONIONS (das Akronym steht für ONtologic Integration Of Naive Sources) ursprünglich um einen synthetischen Ansatz handelt, müssen nach Gangemi et al. die einzelnen Mechanismen, die zu einer Einführung von Begriffen in den bereits existierenden Ontologien geführt haben, einer Prüfung und Integration unterzogen werden. Teilnehmer der Erstellung sind Experten der Wissensrepräsentation, die Fachgebietsexperten werden insofern beteiligt, als dass sie die Plausibilität der Modellierung überprüfen. Eine Kooperation im engeren Sinne findet nicht statt.
Die Ontologieerstellung nach ONIONS verläuft in mehreren Phasen. In der ersten Phase werden mehrere für die Ontologie geeignete Quellen ausgewählt und eine Extraktion der Begriffe und der Ober- und Unterbegriffsanordnung vorgenommen. Die jeweiligen Ordnungen der Begriffe durch eine Relation $\leq$ können dabei nach völlig unterschiedlichen Kriterien zustande gekommen sein. Diese Kriterien und allgemeiner die formalen Kriterien für die Existenz eines Begriffes werden in den nächsten beiden Phasen nachvollzogen, zunächst für jede einzelne Ontologie, dann vergleichend. Es kann sich um induktive oder deduktive Kriterien handeln, intuitive Aspekte kommen nicht zum Tragen, da Gangemi auch nicht von einer informellen Begriffsdefinition ausgeht. Schließlich bildet ein integriertes formales Modell zur Definition der Begriffe in der zu erstellenden Ontologie den Schlusspunkt der Erstellung nach der ONIONS-Methode.

3.4 Die Methontology-Methode

Methontology ist eine Erstellungsmethode nach Gomez-Perez und Fernando-Lopez[36], die unter anderem auf die Erstellung einer Ontologie chemischen Grundlagenwissens angewandt wurde. An der Ontologieerstellung nach Methontology sind sowohl Experten der Wissensrepräsentation als auch Fachexperten des Wissensgebietes der Ontologie beteiligt. Die Aufgabentrennung ist sehr deutlich, denn das Festhalten der Ergebnisse obliegt allein den Experten der Wissensrepräsentation.
Eine Kernforderung, die die Methontology-Gruppe für die Methodik aufstellt, ist die Unabhängigkeit der Ontologieentwicklung von der Zielsprache, ähnlich zu der von uns aufgestellten formalsprachunabhängigen Ontologiedefinition. Als Hauptargument hierfür wird eine Forderung Grubers aufgeführt [38], die bei der Erstellung einer Ontologie eine minimale Störung des Aussagegehalts durch die Anpassung an die Ausdrucksmöglichkeiten einer formalen Sprache verlangt. Dieses Argument ist in zweierlei Weise zu verstehen.

Zum einen können in einer gewählten Sprache Ausdrucksmöglichkeiten fehlen, wie dies beispielsweise bei der Sprache RDF (Ressource Description Framework) [83] in Ermangelung von Subtyp-Funktionalitäten der Fall war und zur Erweiterung durch OWL (Web Ontology Language) [78] geführt hat. Andererseits kann eine Störung des eigentlich beabsichtigten Aussagegehalts insofern erfolgen, als dass entweder die formale Sprache ein Hindernis für die Experten des Wissensgebietes oder eine nur mittelbar über die Experten für Wissensrepräsentation verfügbare Technik darstellt.
In einer Vorbereitungsphase ziehen bei Methontology die Ontologieentwickler Fachliteratur und Gespräche mit den Experten des Fachgebietes zu Rate. Hier erfolgt somit eine Einarbeitung der Ontologieexperten in das Wissensgebiet. Die in dieser Einarbeitungszeit seitens der Ontologieexperten gewonnenen Erkenntnisse fließen in Form allgemeiner Strukturierung intuitiv in die Erstellung ein.
Sodann identifizieren die Ontologieexperten formal umsetzbare Quellen für eine Sammlung relevanter Fachtermini. Diese können Begriffsnamen, Begriffsmerkmale, Begriffsinstanzen, Relationsnamen oder Axiome sein, eine entsprechende Zuordnung erfolgt, zusammen mit der Herstellung von Ober- und Unterbegriffsrelationen, im nächsten Schritt. Die eigentliche erste Verwendung der Relation $\leq$ obliegt wiederum den Ontologieexperten, wobei allerdings mehrere hierarchische Anordnungen der Begriffe arbeitsteilig aufgebaut werden. Das Vorgehen verläuft induktiv, beispielsweise wurden bei der Ontologie zum Fachgebiet Chemie aus Eigenschaften chemischer Elemente (der konkreten Fälle beim induktiven Vorgehen) allgemeinere Begriffe definiert. Die Arbeitsteilung ist nicht als synthetischer Aspekt zu verstehen, vielmehr werden unterschiedliche Bereiche des Wissensgebietes identifiziert und dazu die jeweilige Ober- und Unterbegriffsstruktur festgelegt.
Sobald eine schriftliche und tabellarische Spezifikation eines solchen Modells vorliegt, werden wieder die Fachgebietsexperten befragt und eine Evaluierung und eventuelle Korrektur der Ontologie vorgenommen. Erst im letzten Schritt setzen die Ontologieexperten die schriftliche und tabellarische Ontologiespezifikation mittels der Schnittstelle ODE (Ontology Design Environment) [5] in die Sprache Ontolingua, die als ausdrucksstärkste Ontologiesprache gilt [17], um. Hier können dann auch deduktive Aspekte Berücksichtigung finden, indem aus den vorliegenden Begriffen mit den formalen Mitteln Ontolinguas weitere Begriffe abgeleitet werden. Hinzu kommt die Möglichkeit, gemäß des Integrationsvorschlags von Guarino eine Bereinigung der Ober-Unterbegriffshierarchien aufgrund logischer Metaeigenschaften von Begriffen durchzuführen [33]. Auch dies ist den deduktiven Methoden zuzurechnen.

Kooperativ verläuft die Methontology-Methodik vordringlich beim Austausch der Ontologieexperten und der Fachgebietsexperten. Eine explizite direkte oder durch Ontologieexperten vermittelte Kooperation von Fachgebietsexperten untereinander erklären Fernandez-Lopez und Gomez-Perez jedoch nicht.
Insgesamt gesehen finden wir im Methontology-Verfahren alle Aspekte der Klassifikation nach Holsapple und Joshi wieder, es handelt sich somit wiederum um ein Beispiel dafür, dass die Klassifikation nicht strikt anzuwenden ist.

3.5 Die Delphi-Methode

Die Delphi-Methode baut auf einer schriftlichen Befragung von Experten auf. Dabei handelt es sich um einen systematischen, mehrstufigen Prozess nach Bortz [12]. Die Methode dient dazu, zukünftige Ereignisse, Tendenzen und Entwicklungen gut einschätzen und bewerten können.
Die Befragung nach der Delphi-Methode findet individuell statt, um Gruppengedanken und gemeinsame Denkeffekte zunächst auszuschließen und zu einem möglichst umfassenden Meinungsbild zu gelangen. Bei der Strenge der so vorgenommenen Separierung sind Varianten des Delphi-Modells denkbar [34].
Im Allgemeinen wird eine Gruppe von Experten ein Fragebogen eines gegebenen Wissensgebietes vom so genannten Moderator vorgelegt. Nachdem der Fragebogen beantwortet wurde, werden die schriftlich erhaltenen Antworten, Einschätzungen, Ergebnisse und Meinungen aufgelistet und vom Moderator zusammengefasst. Dieser wertet die Fragebögen aus und entwickelt demnach einen neuen Fragebogen. Eine zweite Befragung folgt mittels des neuen Fragebogens und unter Berücksichtigung der Ergebnisse der ersten Befragung. Dieses Mal sollen die Experten ihre Antworten wechselseitig anonym kommentieren und einen Vergleich zu den Gruppenergebnissen vornehmen. Es folgen weitere Diskussionen, Klärungen und Verfeinerungen der Fragen und Antworten durch Wiederholung dieses Schrittes.
Dieser kontrollierte Prozess wird so oft wiederholt, bis ein zufrieden stellendes Endergebnis, ein allgemein akzeptierter Konsens erzielt wurde. Das vom Moderator erarbeitete Endergebnis ist eine Ideensammlung und Gruppenmeinung, die die Aussagen selbst und Angaben über die Bandbreite vorhandener Meinungen enthält.
Charakteristisch für die Delphi-Methode sind demzufolge folgende Merkmale:

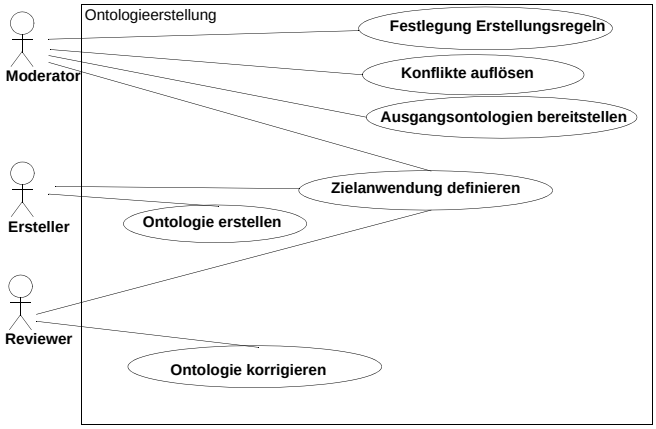

Abbildung 3.1: Aufgabenbereiche innerhalb der Ontologieerstellung

- Festlegung einer Moderatorenrolle
- Verwendung eines formalisierten Fragebogens
- Befragung von Experten
- Anonymität der Einzelantworten
- Ermittlung einer repräsentativen Gruppenantwort
- Information der Teilnehmer über die Gruppenantwort mit strukturierenden Kommentaren des Moderators
- (mehrfache) Wiederholung der Befragung

In der Arbeit von [45] wurde dieses Modell auf die Ontologieerstellung übertragen. Das Anwendungsfalldiagramm in Abbildung 3.1 zeigt die verschiedenen Rollen bei der Ontologieerstellung nach diesem Modell. Dabei fällt auf, dass zwar ein Ontologieexperte als Moderator im Gesamtablauf beteiligt ist, die eigentliche Ontologieerstellung und -korrektur allerdings von Fachexperten vorgenommen wird, wie in Abbildung 3.2 dargestellt.

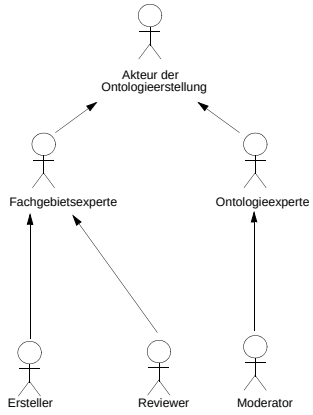

Abbildung 3.2: Rollenverteilung bei der Ontologieerstellung

Eine gemeinsame Aktivität des Moderators und der Fachgebietsexperten stellt die Festlegung von Anwendungen der Ontologie dar. In technischer Hinsicht fällt auch noch die Bereitstellung bereits vorhandener Ontologien aus dem Fachgebiet in den Zuständigkeitsbereich des Moderators.

Holsapple und Joshi spezifizieren den Ablauf ihrer Methodik ähnlich wie in Abbildung 3.3 dargestellt. Der Prozess beginnt mit einer Vorbereitungsphase, in der alle Teilnehmer gemeinsam unter Anleitung des Moderators Designkriterien, Evaluationsstandards und Randbedingungen festlegen. Als Beispiele für Designkriterien können ein induktives oder deduktives Paradigma angesehen werden, so dass auch bei dieser Methode mehrere Aspekte der eingangs dargelegten Klassifikation auftreten können. Randbedingungen beziehen sich hauptsächlich auf das Ziel der zu erstellenden Ontologie und auf ihr Verhältnis zu bereits bestehenden Ontologien, Wissensrepräsentationen oder informelleren Quellen des Fachgebietes. Auch eine Abgrenzung des darzustellenden Wissensgebietes sollte in dieser Phase erfolgen, ebenso wie ein Design für potentielle spätere Erweiterungen der Ontologie.
Aus den in der Vorbereitungsphase identifizierten Quellen oder aber durch erste Eingaben der Fachgebietsexperten bezüglich der Begriffe und ihrer

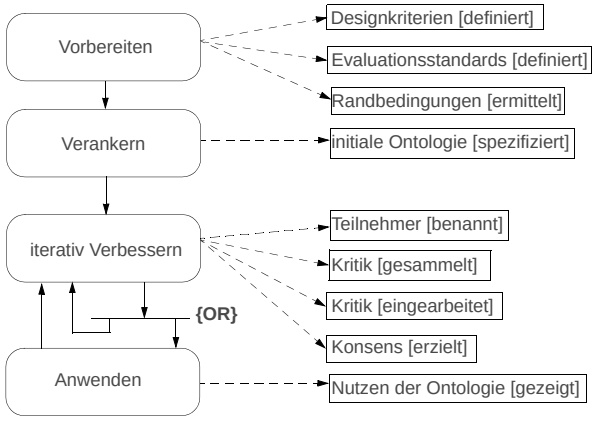

Abbildung 3.3: Erweitertes Delphi-Modell

Ordnung durch $\leq$ stellt der Moderator sodann in der Verankerungsphase eine initiale Ontologie für die Fachgebietsexperten bereit. Diese Spezifikation kann entweder mit Hilfe eines elektronischen Ontologieerstellungswerkzeuges oder in Papierform zur Verfügung gestellt werden.
In der darauf folgende Phase kritisiert ein genau festgelegter Kreis von Fachgebietsexperten die vorhandene initiale Ontologie. Wie Abbildung 3.2 zeigt, können diese Reviewer von den Fachgebietsexperten, die in der Verankerungs- und Vorbereitungsphase für bestimmte Teilbereiche der Ontologie zuständig waren, abweichen.
Holsapple und Joshi schlagen zur Systematisierung der Kritik die Anwendung einer kardinalen Skalierung vor, die eine Form der Benotung von Begriffen vorgibt. Diese Form der systematischen Kritik wurde auch von Schumacher [87] in seiner Arbeit zu Security Patterns (vergleiche Anhang B) angewandt und sollte zusammen mit einer Begründung der numerisch gefassten Kritik durch den einzelnen Fachgebietsexperten erfolgen. Prinzipiell sind hier auch andere Formen der Befragung denkbar. Als zentrale Instanz bei der Sammlung der Kritik sollte jedoch der Moderator fungieren. Unabhängig von der Form der Kritik muss diese vom Moderator in konkrete

Handlungsanweisungen umsetzbar sein. Die Verbesserungsvorschläge werden vom Moderator oder unter Aufsicht des Moderators in die Spezifikation der Ontologie eingearbeitet.
Für das weitere Vorgehen fällt die Darstellung in Abbildung 3.3 allgemeiner aus, als die ursprünglich von Holsapple und Joshi vorgesehene. Im ursprünglichen Delphi-Modell der Ontologieerstellung war vorgesehen, dass nach einer mehrfachen Wiederholung der Kritik und Verbesserung, also einer Iteration der Ontologie, sofort die Anwendung der Ontologie erfolgt. Dieser Weg ist im allgemeineren Fall weiterhin möglich, allerdings können sich, ohne dass das Modell verworfen werden muss, auch während der Anwendung einer Ontologie noch Änderungen und Ergänzungen ergeben. Beispielsweise betrifft dies die Möglichkeit, einen weiteren Sprachgebrauch oder sogar neue Begriffe mit in die Ontologie einzubeziehen. Aus diesem Grund kann ein Wechsel zwischen Anwendungsphase und iterativer Verbesserung erfolgen, solange der Moderator als Instanz, die die Einhaltung der ursprünglichen Designkriterien und Ziele sowie die Auflösung von Konflikten gewährleistet, beibehalten wird.

3.6 Ontologieerstellung in k-med

Der vorliegende Abschnitt beschreibt das Projekt k-med (knowledge based multimedia medical education)[43],[53]. Die Durchführung eines Ontologieerstellungsprozesses in k-med zeigt, welche Art der automatischen Unterstützung für die Ersteller einer Ontologie besonders wichtig sind.
Wir verwenden den Terminus 'Multimedia' nach der Definition von Steinmetz [96], die 'Multimediasysteme' im engeren Sinne erklärt:

Definition 5 (Multimediasystem nach Steinmetz) *'Ein Multimediasystem ist durch die rechnergesteuerte, integrierte Erzeugung, Manipulation, Darstellung, Speicherung und Kommunikation von unabhängigen Informationen gekennzeichnet, die in mindestens einem kontinuierlichen (zeitabhängigen) und einem diskreten (zeitunabhängigen) Medium kodiert sind.'*

Innerhalb des Projektes k-med entstand unter Anleitung des Verfassers der vorliegenden Arbeit eine Ontologie, die zur Beschreibung, Verwaltung und Anzeige multimedialer Unterrichtsmaterialien aus der medizinischen Hochschullehre dient. Darunter fallen - jeweils in elektronischer Form - Texte, Grafiken, Videos, Tondokumente und Animationen. Der Ontologie kommt durch die im Projekt k-med angestrebte Modularisierung des elektronischen Unterrichtsmaterials eine zentrale Bedeutung zu. Verlässt man in

der computerunterstützten Lehre die klassischen Kursstrukturen, die auch bei der Zusammenfügung und Präsentation multimedialer Unterrichtsmaterialien gelten können, so benötigen sowohl die beteiligten Autoren als auch die Zielgruppe der Lernenden eine Orientierung bezüglich des Sinngehaltes der einzelnen Module. Denn die Module können bei abnehmender Größe nur im Idealfall als selbsterklärende Einheiten verwendet werden. Die Beschreibung der Module wird in k-med durch **Metadaten**, die eine festgelegte Beschreibung (durch Felder wir 'Titel', 'Schlüsselwörter' et cetara[89]) jedes einzelnen Moduls darstellen, und durch Ontologien, die als Vokabular für die Schlüsselwörter dienen, geleistet. Einzelne kleinere Informationseinheiten werden durch ihre normierte Beschreibung in einen größeren Sinnzusammenhang gebracht.
Eine modulare Wissensbasis, die die ursprüngliche Architektur von k-med darstellte, wurde durch die Ergebnisse von Seeberg [89] begründet. In diesen Arbeiten zu modularen Wissensbasen stellt Seeberg - über Abschnitt 2.3.1 hinaus gehende - spezifische Vorteile der Einführung von Metadaten und Ontologien heraus. Auch sie bewegt sich dabei im Kontext der computergestützten Lehre. Als Hauptvorteil der Modularisierung im Allgemeinen wird hier die Wiederverwendung multimedialer Unterrichtsressourcen genannt. Dies ist vor allem vor dem Hintergrund hoher Kosten aufwändiger Multimediaproduktion zu sehen. Als besonderen Vorteil einer Modularisierung ist nach Seeberg die inhaltliche Adaptivität bezüglich eines Benutzerprofils zu nennen: damit ist es prinzipiell möglich, bestehende Zusammenstellungen des Unterrichtsmaterials für einen bestimmten Nutzer zu rekombinieren oder sogar automatisch zu erzeugen.
Wir gehen nun auf die einzelnen Phasen der Ontologieerstellung in k-med ein. Danach folgt eine Beurteilung des Erstellungsprozesses, bei der wir die Benutzerumfrage der instruktionswissenschaftlichen Arbeitsgruppe zu Rate ziehen. Wir schließen mit einer Formulierung von Anforderungen, deren Erfüllung zu einer weiteren Erleichterung des Ontologieerstellungsprozesses in verwandten Projekten führen kann.

3.6.1 Eignung der vorgestellten Methoden

Der folgende Abschnitt zeigt auf, welcher Ontologieerstellungsprozess für die Arbeiten in k-med am besten geeignet erscheint.
Die ONIONS-Methode bezieht sich auf einen Erstellungsprozess, der zwar bereits existierende Ontologien aufgreift, den Schwerpunkt jedoch darauf legt, Erklärungen für deren Begriffe und Begriffsstrukturen zu finden. Dies geht weit über den Formalisierungsgrad, der durch Definition 4 gegeben wur-

de hinaus und kann für ein Projekt wie k-med, bei dem eine große Zahl von Anwendern einer nur geringen Anzahl von Experten der Wissensrepräsentation gegenübersteht als zu aufwändig verworfen werden.
Ein ähnliches Argument betrifft Uschold's Methode, da die formale Phase dort stark von der informellen Phase getrennt ist und so Arbeiten, die parallel ablaufen könnten, sequenziell organisiert sind. Zudem erscheint, wenn man wie Gangemi et al. bestehenden Quellen von Ontologien bereits ihre eigene interne Logik a priori zuspricht, die Spezifikation mittels der Kompetenzfragen in Uschold's Ansatz als zu ausgedehnt. Wenn es beispielsweise Gegenstand der Ontologieerstellung ist, wie der Sinngehalt eines tabellarischen oder taxonomischen Standardwerkes aus dem Fachgebiet möglichst genau als Ontologie umgesetzt werden kann, so sollte eine Kompetenzfrage stärker auf dessen Gesamtaufbau abzielen. Zwar ist dies prinzipiell bei Uschold nicht ausgeschlossen, bevorzugt wird jedoch die Entwicklung der Ober- und Unterbegriffsbezüge aus den von den Fachgebietsexperten formulierten Kompetenzfragen.
Die Methontology-Methode stellt dagegen zu stark auf eine strikte Formalisierung und den intensiven Einsatz mehrerer Ontologieexperten ab. Dieses Verfahren ist bei größeren Projekten zu langsam und zu teuer, da die Kritik der Fachgebietsexperten erst sehr spät wieder in die Interaktion mit einbezogen wird.
Aus der Moderation innerhalb der Delphi-Methode ergibt sich hingegen ein großer Vorteil für die Befragten, in diesem Falle die Mediziner als Fachgebietsexperten. Vor allem nach der ersten Rückmeldung können sie ihre Meinung ändern, ohne in unproduktive fachliche Konkurrenzsituationen zu geraten. Ein weiterer Vorteil der Delphi-Methode ist, dass die Ontologieerstellung nicht durch Konformitätsstreben und durch die Dominanz einer oder mehrerer Personen in einer Gruppe beeinträchtigt wird. Diese Vorteile gelten auch dann, wenn wie in k-med ohne eine kardinale Bewertung oder fest vorgegebene schriftliche Fragebögen gearbeitet wird.
Als Nachteil der Delphi-Methode kommt der Zeitaufwand für das Durchführen der Befragungen in Betracht. Hinzu kommt, dass bei der deutlichen fachlichen Unterteilung, wie sie in k-med vorliegt, die vollständige Anonymität der beteiligten Experten nicht immer gewahrt bleiben kann und stattdessen wiederum die Rolle des Moderators als Mittler der Meinungen gefragt ist. Die Nachteile werden jedoch durch eine genau definierte Begleitung des Erstellungsprozesses ausgeglichen und zeitliche und inhaltliche Verluste können - bei geeigneter Wahl von Ontologieerstellungswerkzeugen - durch die im Gegensatz zu den anderen vorgestellten Methoden direkte Arbeit der Fachgebietsexperten mit der Wissensrepräsentation ausgeglichen werden. Dies

ist dann umsetzbar, wenn die Kritik innerhalb eines vom Moderator festgelegten Rahmens direkt zu Änderungen oder Erweiterungen der Ontologie führen darf oder wenn die vom Moderator gebündelte Kritik von mehreren Fachgebietsexperten parallel eingearbeitet werden kann. Insgesamt werden die Fachgebietsexperten stärker in die eigentliche Arbeit eingebunden, begünstigt durch einen Formalisierungsgrad, wie er im zweiten Kapitel der vorliegenden Arbeit begründet wurde und wie er in der Standardvisualisierung aus Abbildung 2.4 auch Laien der Wissensrepräsentation zugänglich ist. Die explizite Erwähnung bestehender Quellen für die initiale Ontologie und der ergebnisoffene Umgang mit der initialen Ontologie bilden bei der kooperativen Ontologieerstellung eine ideale, den jeweiligen Projektzielen anpassbare Verallgemeinerung der Umgangsweisen mit bestehenden Quellen, wie Uschold, Fernandez-Perez et. al. sowie Gangemi sie vorschlagen.
Für die Arbeiten im k-med Projekt wurde die Methode nach Holsapple und Joshi gewählt und in einigen Einzelheiten weiter interpretiert. Wir gehen im Rest des vorliegenden Kapitels auf die Anwendung der Methode in k-med ein und schließen mit Anforderungen zu einer weiteren Vereinfachung des kooperativen Erstellungsprozesses, die aus den Arbeiten im Projekt abzuleiten sind.

3.6.2 Anwendung des Delphi-Modells

Der folgende Abschnitt zeigt, wie das Delphi-Modell in k-med angewandt wurde. Anhand dieser Anwendung werden die für das Gesamtziel der vorliegenden Arbeit wichtigen Anforderungen an eine ähnlichkeitsbasierte automatische Unterstützung der Ontologieerstellung plausibel.
Die Ontologie in k-med besteht aus medizinischen Fachbegriffen. Sie wurde nach einem leicht modifizierten Delphi-Modell erstellt. Zentral war hierbei ein abgestimmtes Vorgehen der medizinischen Autoren im Projekt. Eine Besonderheit von k-med ist, dass mehrere medizinische Fächer des Grundstudiums vertreten sind. Dabei können auch pro Fach mehrere Autoren, die sich an verschiedenen Orten befinden, vertreten sein.

3.6.2.1 Vorbereitungsphase

Die Abstimmung der Autoren betrifft die Festlegung eines gemeinsamen Vokabulars (der Namen der Elemente aus B gemäß Definition 4 und ihrer Zusatzannahmen) , die Festlegung der Relationen (der Menge R gemäß Definition 4) und darüber hinaus den Detaillierungsgrad der Ontologie, wie er auch schon in der ONIONS-Methodik als 'stopover problem' identifiziert

wurde. Weiterer Abstimmungsbedarf entsteht bei der Auflösung von Konfliktfällen bei sich überschneidenden Wissensgebieten. Dies ist im Projekt k-med von divergierenden Auffassungen innerhalb eines medizinischen Fachgebietes zu unterscheiden. Schließlich ist die Rolle des Moderators zu klären, da er in k-med eine Zusatzfunktion als Schulungsleiter für die Bedienung der Ontologieerstellungswerkzeuge innehatte.
Wir beziehen uns nun im einzelnen auf die in der folgenden Übersichtstabelle 3.1 dargestellten typischen Phasen nach dem Delphi-Modell.

Tabelle 3.1: Phasen der Ontologieerstellung in k-med

Phase	*Ergebnisse*
Vorbereitung	Definition von Designkriterien: Quelle, Detaillierungsgrad, Relationen Ermittlung von Randbedingungen: Moderator, Evaluationsstandards
Verankerung	Spezifikation der initialen Ontologie aus den Katalogen Überschneidung der Fachgebiete
iterative Verbesserung	(stufenweise) Modifikation der initialen Ontologie
Anwendung	systematische Anfragen, Suche, Wiederverwendung von Modulen

Einerseits liegt eine strikte Abfolge der Phasen Vorbereitung, Verankerung und dem Beginn der iterativen Verbesserung vor, andererseits findet nach der Verankerungsphase ein steter Wechsel zwischen iterativer Verbesserung und Anwendung der Ontologie statt[1]. Hier ist eine leichte Modifikation des Ansatzes von Holsapple und Joshi zu erkennen, die aus der Anwendung der Ontologie als Vokabular für die Metadaten resultiert. Wir werden bei der Darstellung der einzelnen Phasen näher auf besagte Wechselwirkung eingehen.
Als Ergebnisse der der Vorbereitungsphase ist festzuhalten: der so genannte Gegenstandskatalog der beteiligten Fächer Physiologie, Histologie, Pharmakologie, Dermatologie, Nuklearmedizin und Biochemie. Diese Kataloge

[1]Dies geht bei k-med damit einher, dass die Ersteller der Ontologie gleichzeitig auch Anwender der Ontologie sind. Die Verschlagwortung von Unterrichtsmaterial ist eine der Anwendungen der Ontologie in k-med.

sind allgemein verfügbar [47]. Für die ebenfalls beteiligte Infektiologie lag kein Gegenstandskatalog vor, so dass hier eine besondere Lösung gefunden werden musste. Die Infektiologie wurde durch das Inhaltsverzeichnis eines Lehrbuches in die initiale Ontologie aufgenommen.
Gegenstandskataloge existieren für alle vorklinischen Fächer des Medizinstudiums. Sie umfassen die von den Hochschullehrern verbindlich zu unterrichtenden Themengebiete. Vorteilhaft an einer Fundierung der Ontologie durch die Gegenstandskataloge ist der bereits vorhandene Bezug der Medizinstudierenden zum in ihnen festgehaltenen Vokabular. Der Gegenstandskatalog dient in der Regel als Anhaltspunkt für die Stoffmenge bei der Vorbereitung auf die Physikumsprüfungen. Es folgt ein Ausschnitt aus dem Gegenstandskatalog der Biochemie:

Chemie für Mediziner und Biochemie (Inhaltsübersicht)
GRUNDLAGEN DER CHEMIE
1.1 Makroskopische Erscheinungsformen der Materie
2 Aufbau und Eigenschaften der Materie 2.1 Atome, Isotope, Periodensystem 2.2 Chemische Bindung 2.3 Acyclische Kohlenstoffverbindungen, einfache funktionelle Gruppen 2.4 Carbo- und Heterocyclen 2.5 Stereochemie
3 Stoffumwandlungen 3.1 Homogene Gleichgewichtsreaktionen 3.2 Heterogene Gleichgewichtsreaktionen 3.3 Säure/Base-Reaktionen 3.4 Redox-Reaktionen 3.5 Bildung und Eigenschaften der Salze 3.6 Ligandenaustausch-Reaktionen 3.7 Additions/Eliminierungs-Reaktionen 3.8 Substitutionsreaktionen 3.9 Sonstige Reaktionen
CHEMIE BIOLOGISCH UND MEDIZINISCH RELEVANTER NATURSTOFFE
4 Kohlenhydrate 4.1 Monosaccharide 4.2 Disaccharide 4.3 Oligo- und Polysaccharide 5 Aminosäuren, Peptide, Proteine 5.1 Aminosäuren 5.2 Peptide 5.3 Proteine 6 Fettsäuren, Lipide 6.1 Fettsäuren 6.2 Acylglycerine 6.3 Sphingolipide 6.4 Steroide 7 Nukleotide, Nukleinsäuren, Chromatin 7.1 Nukleotide 7.2 Nukleinsäuren 7.3 Chromatin 8 Vitamine, Vitaminderivate, Coenzyme 8.1 Allgemeines 8.2 Biochemischer Mechanismus 8.3 Pathobiochemie

Die Nummerierung und die Hauptüberschriften können in eine hierarchische Struktur überführt werden, nämlich die der Unterbegriffsrelation $\leq$ gemäß Definition 4. Beispielsweise entstehen so aus den Gliederungspunkten 4 und 5 unter anderem die Beziehungen

$$\textit{Monosaccharide} \leq \textit{Kohlenhydrate} \leq \textit{Chemie biologisch und medizinisch relevanter Naturstoffe}$$

sowie

Pathobiochemie $\leq$ *Vitamine, Vitaminderivate, Coenzyme* $\leq$ *Chemie biologisch und medizinisch relevanter Naturstoffe*

Dem Verständnis leichtgewichtiger Ontologien entsprechend, sollte besonders im zweiten Fall der Begriffsbezeichner *Pathobiochemie* als Begriff *Pathobiochemie als Thema des medizinischen Lehrplans* verstanden werden. Dieses Versändnis ist zur Akzeptanz der hier vorliegenden Relation $\leq$ nötig. Im Gegenstandskatalog existieren zu den zweistellig nummerierten schlagwortartigen Überschriften, die wir oben aufgelistet haben, auch noch weitere spezielle Stichwortlisten, die zur Erläuterung eines nummerierten Begriffes dienen. Für den Detaillierungsgrad der Begriffe in der k-med-Ontologie wurde in der Vorbereitungsphase vereinbart, dass nur dann zusätzliche Begriffe zu den ersten drei Hierarchieebenen aufgenommen werden, wenn sie nötig sind, um k-med Lernobjekte zu beschreiben. Umgekehrt wird beispielsweise aus Gründen der Übersichtlichkeit davon abgeraten, ohne ein im k-med System existierendes Lernobjekt zum Thema '*Vitamin B12*' den Begriff '*Vitamin B12*' als Unterbegriff zu *Vitamine, Vitaminderivate, Coenzyme.* einzuführen. Als Ganzes gesehen entsteht damit schon in der Vorbereitungsphase für die spätere Anwendungsphase die Vereinbarung, dass pro Lernobjekt mindestens ein (und idealerweise genau ein) Begriff in der Ontologie vorhanden sein muss.
Wie in der medizinischen Terminologie und auch in den Gegenstandskatalogen üblich, verwenden wir den Nominativ der Pluralform eines einzelnen Fachbegriffes, sofern diese Pluralform existiert. Dies ist im obigen Beispiel bei *Vitamine, Vitaminderivate, Coenzyme* der Fall. Sofern die Pluralform nicht existiert, verwenden wir den Nominativ Singular. Dies ist etwa bei *Chemie biologisch und medizinisch relevanter Naturstoffe* der Fall.
Ebenfalls Gegenstand der Vorbereitungsphase in k-med war eine Anforderungsanalyse bezüglich der Darstellungsform der Ontologie für den Anwender. Diese Analyse hat mehrere Aspekte. Bei der Visualisierung der Ontologie für den Erstellungsprozess bestand die Alternative zwischen der Darstellung, wie wir sie bislang in den Grafiken der vorliegenden Arbeit gewählt haben und einer von Steinacker [95] vorgeschlagenen Variante. Letztere nutzt stärker die gesamte beim grafischen Erstellungsprozess zur Verfügung stehende Fläche, dabei fallen allerdings die hierarchischen Strukturen in der Form, dass abstraktere Begriffe in den oberen Bereichen und speziellere Begriffe in den unteren Bereichen des Bildschirms angesiedelt werden, weg. Die medizinischen Autoren bevorzugten klar eine Beibehaltung dieses Layoutgrundsatzes. Die angestrebte Visualisierung, umfangreiche lexikalische Suchfunktionen beim Erstellen und Pflegen einer Ontologie sowie

weit reichende kooperative Funktionen waren am besten in den k-infinity-Erstellungswerkzeugen der Firma intelligent views[48] verwirklicht, die für die Konstruktion der k-med-Ontologie eingesetzt werden konnten.
Ein weiterer Aspekt war die Vorauswahl der Quellen für die Ontologie. Durch die Auswahl der Gegenstandskataloge wird nach übereinstimmender Meinung der beteiligten Mediziner eine geringe Verzerrung bei der Kodierung erreicht. Andere bestehende medizinische Ontologien wie GALEN [76] erforderten nicht nur eine eingehende formale Ausbildung der beteiligten Autoren, sondern sind für ein Projekt wie k-med zu stark an den eigenen Darstellungsformalismus gekoppelt. Im Gegensatz dazu ist den Medizinern der hierarchische Aufbau der Gegenstandskataloge bereits vertraut. Aus ähnlichen Gründen wurde eine Erstellung der Ontologie mit Hilfe des UMLS Metathesaurus [74] verworfen. Während bei den verwendeten Gegenstandskatalogen eine eindeutige Zuordnung nach den beteiligten Fachgebieten möglich ist, entstünde bei UMLS das Problem, wie die Arbeit an den einzelnen Bereichen der entstehenden Ontologie den Beteiligten zuzuordnen sei. Diese Frage könnte nur in einem zusätzlichen Arbeitsschritt beantwortet werden, was angesichts der mehrfachen Arbeitsbelastung der Autoren vermieden werden sollte.
Eine erste Absprache über Relationsnamen wurde ebenfalls in der Vorbereitungsphase getroffen. Da gleichzeitig Übereinstimmung darüber herrschte, dass die ursprüngliche Liste der Relationsnamen nochmals zu überprüfen war, sobald die initiale Ontologie zur Verfügung stehen sollte, erfolgt eine detaillierte Aufstellung weiter unten. Hier sei lediglich erwähnt, dass zur besseren Orientierung jede Relation entweder symmetrisch angelegt sein sollte oder eine benannte Umkehrrelation besitzen sollte. In der Sprechweise der Definition 4 heißt das, wenn zwei Begriffe a und b in der Relation r_i zueinander stehen:

$$\forall r_i \in R \exists {r_i}^{-1} \in R : r_i(a,b) \Leftrightarrow {r_i}^{-1}(b,a) \tag{3.3}$$

Allgemein wurde zu der Anwendung der Relationen noch vereinbart, dass nur der Moderator neue Relationsnamen erstellen darf, so dass eine willkürliche Explosion vorhandener, an sich aber sehr ähnlicher Erstellungsmöglichkeiten auszuschließen ist.

3.6.2.2 Verankerungsphase

In der Verankerungsphase wird für die k-med Autoren der Gegenstandskatalog nebst seiner Hierarchie elektronisch umgesetzt und zur Verfügung gestellt. Diese initiale Ontologie steht mittels der k-infinity Werkzeuge der

intelligent views GmbH [48] den Autoren auf zwei Arten zur Verfügung. Die Abbildung 3.4 zeigt eine expansive Sicht der Ontologie mittels einer hierarchisch angeordneten graphenbasierten Darstellung und die alphabetisch geordnete Liste aller initialen Begriffe. Die Schulungen der Fachautoren zur Ontologieerstellung vermittelten die Manipulation des initialen Begriffsnetzes durch Benutzerinteraktionen mit diesen Sichten.
Aus Abbildung 3.4 ist auch die alphabetische Sicht auf die Begriffe zu entnehmen. Es bestehen in diesem Teil der k-infinity Werkzeuge Möglichkeiten der Suche durch die Eingabe von Zeichenketten. Die Zeichenketten können dabei auch geringe Abweichungen von eigentlichen Begriffsnamen enthalten oder nur aus Teilen des eigentlichen Begriffsnamens bestehen. Diese Funktion ist besonders bei ersten Korrekturen des Netzes hilfreich, bei denen Begriffe aus der initialen Ontologie mit ähnlich lautendem Namen identifiziert werden sollten. Zudem bieten die in der Abbildung 3.4 gezeigten, mit Buchstaben gekennzeichneten Schaltflächen die Möglichkeit, zu einem bestimmten Buchstaben alle Begriffe mit entsprechendem Namensanfang in einer Liste aufzurufen.
Insgesamt waren somit reichhaltige Möglichkeiten der Ansicht und Korrektur der initialen Ontologie gegeben, die in der Verankerungsphase genutzt wurden. Die Korrektur umfasste dabei lediglich das vorhandene Vokabular, Erweiterungen der Ontologie sind erst Teil der Wechselwirkung aus den späteren Phasen der iterativen Verbesserung und Anwendung.
Die Autoren erhielten somit zunächst den Auftrag, den generellen Aufbau des umgesetzten Gegenstandskataloges nachzuvollziehen und zu kommentieren. Dies geschah im Rahmen einer Einzelschulung, bei der auch das Softwarewerkzeug zum Betrachten der initialen Ontologie erklärt wurde. Die Überprüfung behandelte auch (neben dem allgemeinen Auftrag, Rechtschreibfehler zu beheben) die korrekte medizinische Terminologie und Schreibweise. Wenn verschiedene Quellen in Form der Gegenstandskataloge zusammengeführt werden, entstehen darüber hinaus Überschneidungen der Begriffsmengen. Wenn wir beispielsweise mit B_p die Menge der Begriffe aus dem Gegenstandskatalog der Physiologie und mit B_b die Menge der Begriffe aus dem Gegenstandskatalog der Biochemie bezeichnen, so gilt

$$B_p \cap B_b \neq \emptyset \tag{3.4}$$

Verwenden zwei oder mehrere Autoren aus unterschiedlichen Fachgebieten den gleichen Begriff b, so müssen diese Autoren in der Verankerungsphase eine Vereinbarung über den weiteren Umgang mir diesem Begriff treffen. Prinzipiell wurden drei verschiedene Lösungsansätze für solche Konfliktfälle verwendet. Zum einen konnten sich die Autoren darauf einigen, dass der

Begriff b gemeinsam genutzt werden sollte, weil er tatsächlich die gleiche Entität beschreibt. In diesem Fall gälten als Ergebnis (für das obige Beispiel der Fachgebiete Physiologie und Biochemie) dann für ein Paar $b_p \in B_p$ und $b_b \in B_b$ mit $b_p, b_b \notin B_p \cap B_b$ die Unterbegriffsbeziehungen

$$b \leq b_p \leq \mathit{Physiologie} \tag{3.5}$$

und

$$b \leq b_q \leq \mathit{Biochemie} \tag{3.6}$$

Die weiteren Möglichkeiten ergaben sich aus der Umbenennung des Begriffes. Diese konnte dadurch entstehen, dass von einem oder von beiden medizinischen Autoren ein speziellerer Name aus dem jeweiligen Fachgebiet gewählt wurde. Wenn ein solcher Name nicht identifizierbar war, so konnte der unterschiedliche Sinngehalt durch einen Namenszusatz, der das jeweilige Fachgebiet angibt, verdeutlicht werden. Bei dieser Lösungsvariante wurde beispielsweise der von der Biochemie und der Physiologie unterschiedlich verwendete Begriff *Absorption* in die Begriffe *Absorption(Physiologie)* und *Absorption(Biochemie)* überführt. Die Autoren wurden mit Hilfe einer tabellarischen Aufstellung der Begriffe der initialen Ontologie auf die Überschneidungen der Fachgebiete aufmerksam gemacht. Durch diese Aufstellung war auch von Seiten des Moderators festgelegt, welche Interaktionen zwischen einzelnen Autoren bei der Auflösung von Konfliktfällen notwendig waren.
Die Liste der Relationen R wurde ebenfalls in der Verankerungsphase aufgestellt. Hierbei wurden zu ähnlich lautende Relationsnamen vermieden. Im Einzelnen lauten die Relationsnamen, die ohne Einschränkung von allen Autoren auch über Fachgebietsgrenzen hinweg verwendet werden durften:

aktiviert, hat Bezug zu, spaltet ab, gehört zu, wird angewendet bei, inaktiviert, beeinflusst, reagiert mit, interagiert mit, ist die Grundlage von, hat Funktion, katalysiert, wird behandelt mit, definiert, dissoziiert zu, interagiert mit, ist Teil von, reagiert zu, ist Test für, assoziiert mit, ist korreliert, ist Ziel von, wirkt auf, unterstützt, wird aktiv gegen, verursacht

Die Relationsnamen zeigen, dass auch hier wieder die Intention der Ontologie in k-med eine entscheidende Rolle spielt. Während Relationen wie *aktiviert* einen fachlichen Charakter aufweisen, sind andere Relationen wie *hat Bezug zu* didaktischer Natur. Dies ist eine Konsequenz des Gesamtszenarios in k-med, welches das medizinische Wissen im Unterrichtskontext

und für die multimediale Lehre erfassen soll.
Die Umkehrrelationen zu den aufgelisteten Relationen orientieren sich weitestgehend am sprachlichen Passiv, wie etwa bei der Relation *wird aktiviert durch*. Bei einzelnen Relationen ist dies nicht möglich, was Relationsnamen wie *hat zur Grundlage* notwendig werden lässt. Einzelne Relationen wie *hat Bezug zu* und *interagiert mit* sind symmetrisch. Ihre Umkehrrelation trägt den gleichen Namen.

3.6.2.3 Iterative Verbesserung und Anwendung

In der Phase der iterativen Verbesserung wird die initiale Ontologie erweitert und gepflegt. Dies obliegt wiederum den Fachautoren unter Anleitung des Moderators. Erweiterungen der Ontologie bestehen aus spezielleren Begriffen gemäß der Verabredungen zum Detaillierungsgrad, sowie aus der Anwendung der zusätzlichen Relationen auf die zunächst nur hierarchisch gegliederte Ontologie. Die Herstellung von σ nach Definition 4 verläuft frei nach dem Dafürhalten der Ersteller. Konfliktfälle können hier dann auftreten, wenn mindestens einer der durch eine Relation verbundenen Begriffe nicht zum Aufgabenbereich desjenigen, der die Relation angelegt hat, gehört. Hier kann eine direkte Absprache Abhilfe schaffen, im Zweifelsfall wird der Moderator hinzugezogen.
Bei der iterativen Verbesserung der Begriffe sind wiederum Überschneidungen der einzelnen Fachgebiete zu beachten, die Auflösung von Konfliktfällen verläuft wie in der Verankerungsphase.

Das Aktivitätsdiagramm in Abbildung 3.5 zeigt an, dass bei der Beschreibung eines Moduls mit einem Begriff aus der Ontologie zwei Fälle auftreten können. Im Idealfall findet der k-med-Autor zu einem Modul, das soeben fertig gestellt wurde, einen geeigneten Begriff in der Ontologie vor, folgt also dem rechten Ast des Diagramms. Die Suche nach dem passenden Begriff läuft innerhalb der Ontologie mit den k-infinity-Werkzeugen ab. In der Praxis bedeutet dies, dass der Autor die Ontologie daraufhin überprüft, ob und wie sie seine gedankliche Vorstellung eines zum Modul passenden Begriffes enthält.

Von zentraler Bedeutung für die Beschreibung eines Moduls in k-med ist die folgende Vereinbarung: Es ist in k-Med ausgeschlossen, einen Begriff in das Metadatenfeld 'Schlüsselwörter' einzutragen, der nicht in der Ontologie enthalten ist. Positiv formuliert bedeutet das: Jede Verschlagwortung eines Moduls in k-med muss durch einen Begriff erfolgen, der in der Ontologie angelegt wurde. Bei einer konsequenten Anwendung der Faustregel ergeben sich zwei Szenarien: der Begriff ist bereits vorhanden oder der Begriff muss,

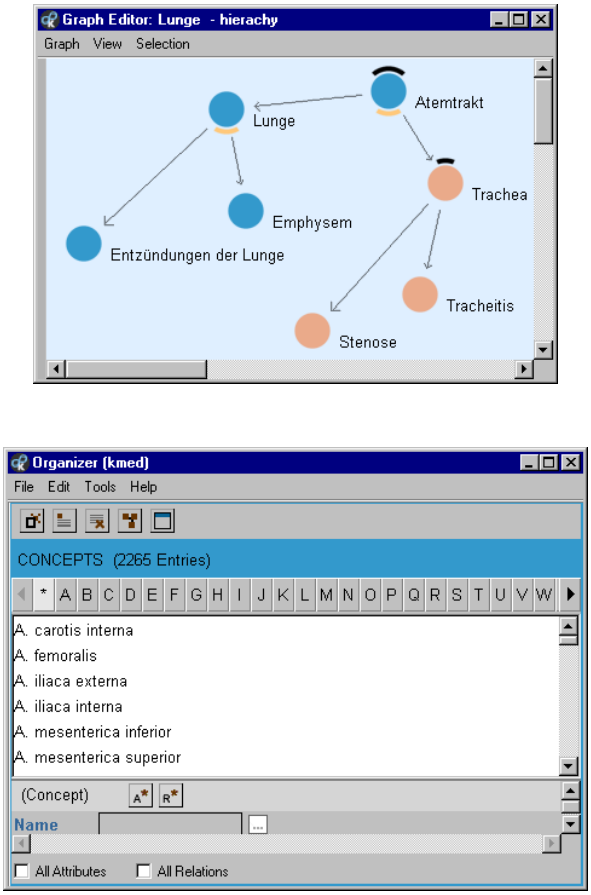

Abbildung 3.4: Hierarchie und Auflistung von k-med Begriffen in der Visualisierung durch k-infinity

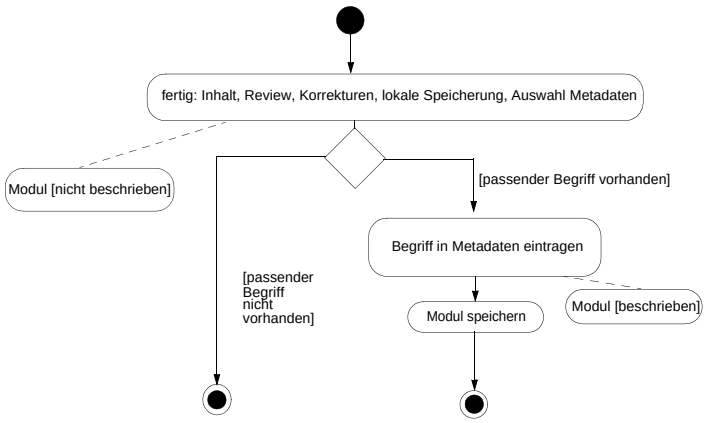

Abbildung 3.5: Iterative Verbesserung bei der Modulerstellung

obschon das Modul selbst bereits existiert, noch angelegt werden. Bei einer konsequenten Anwendung dieser Faustregel ergeben sich die Szenarien, die wir in Abbildung 3.6 darstellen und im Folgenden erläutern.
Das Aktivitätsdiagramm 3.6 zeigt den Ablauf einer Nutzung und Ergänzung der Ontologie in dem Moment, in welchem ein medizinischer Fachbegriff zur Beschreibung eines Moduls mittels des Metadatenfeldes 'Schlüsselwörter' gebraucht wird. Die Diagramme 3.5 und 3.6 wurden getrennt dargestellt, weil leichte Abweichungen in diesem Arbeitsablauf denkbar sind. So erscheint es ebenso sinnvoll, bereits bei der Planung neuer Unterrichtsinhalte die k-med-Ontologie zu ergänzen, wenn Sicherheit darüber besteht, dass auch tatsächlich die entsprechenden Lernmodule erstellt werden. Wenden wir uns nun der Bearbeitung der Ontologie zu.
Zunächst muss, um zum Ausgangspunkt der Abbildung 3.6 zu gelangen, das Erstellungswerkzeug geöffnet werden, um den aktuellen Stand der Ontologie aufzurufen. Daraufhin erfolgt eine Suche nach dem zum vorliegenden, soeben fertig gestellten Modul passenden Begriff b. Hier liegt der passendste Zeitpunkt für eine gedankliche Festlegung des Begriffs b. Mindestanforderung sollte aber eine gedankliche Festlegung von b sein.

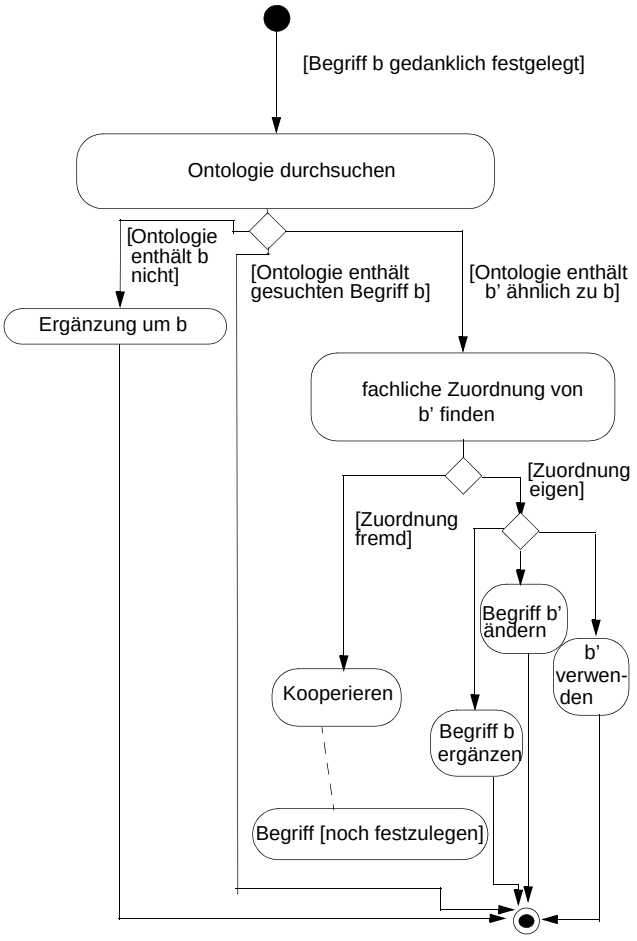

Abbildung 3.6: Fallunterscheidung bei Nutzung oder Neuerstellung von Begriffen in der Anwendungsphase

Dies kann entweder vollzogen werden, indem man rein graphisch vom Oberbegriff des eigenen Faches (*Biochemie*, *Nuklearmedizin*, *Infektiologie*, *Dermatologie*, *Physiologie* oder *Histologie*) in die Spezialgebiete navigiert, wobei man schrittweise den am besten passenden Unterbegriff auswählt. Wir benennen für den Rest dieses Abschnittes den Oberbegriff des gesuchten b mit b_0. Das Verfahren der sukzessiven Navigation unterstellt allerdings eine Vollständigkeit der vorliegenden Begriffe. Eine andere Möglichkeit besteht darin, als Autor vor der eigentlichen Navigation gedanklich einen das Modul beschreibenden Begriff festzulegen. Um zu ermitteln, ob dieser vorbestimmte Begriff auch tatsächlich in seiner exakten Schreibweise in der k-med Ontologie angelegt ist, stellt das Werkzeug immer eine Suche nach orthographisch ähnlichen Begriffen zur Verfügung. Statt eines gesuchten Eintrags *Salmonellen* ist somit ersichtlich, ob in Wirklichkeit *Salmonella* in der Ontologie angelegt wurde und auch als Eintrag für das Metadatenfeld 'Schlüsselbegriffe' verwendet werden sollte.
Es erwies sich als sehr nützlich, mittels der graphischen Exploration auch noch die Umgebung des ausgewählten Begriffes dahingehend zu untersuchen, ob eventuell ein noch speziellerer, besser geeigneter Begriff vorliegt und verwendet werden kann.
Als Ergebnis der Suche nach dem gedanklich oder real festgelegten Begriff b können verschiedene Fälle auftreten.
Der einfachste Fall ist die exakte Übereinstimmung des Suchergebnisses mit b, so dass b direkt verwendet werden könnte, was im mittleren Teil des Aktivitätsdiagramms 3.6 dargestellt wird.
Direkter gestaltet sich die Vorgehensweise, wenn die Ontologie einen Begriff b' enthält, der dem gesuchten b ähnlich ist und für den gleichzeitig

$$\text{[Zuordnung eigen]} :\Leftrightarrow (b' \leq b_0) \tag{3.7}$$

gilt. Hier ergibt sich für den Autoren die freie Wahl, ob der ähnliche Begriff übernommen werden soll, ob er geändert werden soll, oder wie die Ergänzung und relationale Verbindung des zusätzlichen b erfolgen soll.
Wenn der Begriff b noch nicht in der Ontologie vorhanden ist, so sollte er als Unterbegriff eines bestehenden Begriffes ergänzt werden. Die Darstellung auf der linken Seite des Aktivitätsdiagramms 3.6 erscheint hier leicht vereinfacht, da für diesen Fall ein linearer Ablauf angenommen wird. Die Ergänzung des Begriffes b wird derart ablaufen, wenn es sich um einen Begriff aus dem eigenen Fachgebiet handelt, wenn also $b \leq b_0$ gilt. In dem seltenen Fall, dass ein spezieller, in der initialen Ontologie fachfremder Begriff $b \leq b_f$ (wobei $b_f \notin \{b_i \in B | b_i \leq b_0\}$) ergänzt werden müsste, sollte hier eine Abstimmung zwischen dem bei der Ergänzung von b aktiven Autor und

dem Autoren, dem fachlich der Begriff b_f zuzuordnen ist, erfolgen.
Den Fall der Kooperation haben wir explizit in das Diagramm 3.6 eingetragen, wenn ein zu b ähnliches b' gefunden wurde, für das in Analogie zu (2.5)

$$\text{[Zuordnung fremd]} :\Leftrightarrow (b' \leq b_f, b_f \notin \{b_i \in B | b \leq b_0\}) \tag{3.8}$$

gilt. Während der Kooperation, das heißt der Absprache, ob bestehende Begriffe geändert oder ergänzt werden, gibt es für den Autoren, von dem die im Diagramm wiedergegebene Aktion ausging, noch keinen vollständig zu seiner gedanklichen Vorstellung von b passenden Begriff innerhalb der Ontologie. Dieser Fall kann bei zu b ähnlichen Suchergebnissen wesentlich häufiger auftreten als bei identischen Suchergebnissen, da in der Verankerungsphase bereits die Überschneidungen der einzelnen Begriffsmengen pro Autor festgestellt und behoben wurden und somit die Fälle ähnlich lautender oder in ihrer Spezialisierung noch nicht spezifizierter Fachbegriffe übrig bleiben.
Eine Abweichung der Vorgehensweise von der Methodik nach Holsapple und Joshi liegt darin, dass bei der k-med-Ontologie nicht unmittelbar eine Definition der Begriffe vorgenommen wurde. Trotzdem soll auf diese sinnvolle Ergänzung der Ontologie nicht verzichtet werden. Wie wir in unseren Vorbetrachtungen erläutert haben, kann in einem Projekt wie k-med lediglich eine natürlichsprachliche und keine formale Beschreibung der einzelnen Begriffe erfolgen.
Zur Anwendungsphase gehört allerdings ein Spezifikum, das in dieser Form in Projekten aus der elektronischen Lehre, in Dokumentationen und ontologiebasierten Nachschlagewerken entstehen kann, nämlich das Bedürfnis nach einem Glossar. Neben der Auszeichnung von Modulen durch Metadaten stellt dies eine weitere unmittelbare Anwendung der k-med Ontologie dar. Bei einem Glossar wird zu Begriffen eine natürlichsprachliche Erklärung angelegt, somit kann die von Holsapple und Joshi verlangte Spezifikation der Begriffe gleichzeitig mit der Glossarerstellung verbunden werden. Die Funktionalität, die sich hier bietet, geht über die in Glossaren übliche alphabetische Aufstellung einer Anzahl von Fachbegriffen hinaus, da auch ihre Anordnung mittels der Unterbegriffsrelation $\leq$ und ihre Verbindung durch die Relationen aus R gemäß Definition 4 visualisiert werden können. Diese zweifache Verwendung der einzelnen Begriffsdefinitionen setzt allerdings voraus, dass Begriffe auch extern, das heißt außerhalb des Erstellungswerkzeuges in ihrem Zusammehnhang angezeigt werden können. Im Projekt k-med existieren dazu als prinzipielle Möglichkeiten eine Lernplattform und eine WWW-Ansicht der mit k-infinity Werkzeugen erstellten Demonstration

des ontologiebasierten Glossars [52]. Die Speicherung der Glossareinträge im ASCII-Textformat erfolgt in der Datenbank des k-infinity-Systems. Für unsere Betrachtungen ist es lediglich wichtig, dass dort jeder Begriff und auch jeder Glossareintrag mit einem numerischen Identifikator versehen wird.
Die Anbindung an einen k-med Kurs, das heißt, an eine didaktische Zusammenstellung von Modulen, kann über die Lernplattform, die eine http-Anfrage an die Webschnittstelle der k-infinity Wekzeuge absetzt, erfolgen. Auf die http-Anfrage hin wird der interne numerische Identifikator des gesuchten Begriffes zurückgeliefert, wonach ein zweiter http-Request dann nach dem eigentlichen Begriff, seinen fachlichen Relationen zu weiteren Begriffen, den dazugehörigen Glossareinträgen und Literaturhinweisen abgesetzt wird. Schlussendlich werden in Form von XML-Fragmenten Begriffe, Relationen und Identifikatoren zurückgeliefert. Die Anfragen für die Identifikatoren gestalten sich folgendermaßen:

```
http://bridge.host.domainkennung:3030/mb?request=search
&search=handmuskeln&volume=kmed
```

Der erste Teil dient der Erkennung des k-infinity-Servers, `volume` gibt die Ontologie an. Die Interpretation der Zeichenkette `handmuskeln` ist hier eindeutig. Bei einer Suche nach `hand` würden mehrere Identifikatoren, nämlich in unserem Falle diejenigen der Begriffe *Hand* und *Handmuskeln*, in der Antwort erscheinen.
Zurück zu unserem Beispiel: wurde dann der Identifikator (hier: 196375542293) des Begriffes *Handmuskeln* vom k-infinity-Server zurückgeliefert, so erfolgt eine zweite Anfrage:

```
http://bridge.host.domainkennung:3030/ mb?rquest=id42&
const=ID196375542293,bfs,1,0,false&volume=kmed
&docparts=text,title
```

Hier werden neben dem Identifikator und den Namen der Ontologie und des Servers Suchoptionen übergeben, die festlegen, ob auch noch Relationen oder eine weitere Umgebung des Begriffes angezeigt werden soll. Zudem können die Glossareinträge in Text und Titel untergliedert werden, worauf der letzte Teil der Anfrage verweist. Hier soll beides zurückgeliefert werden.
Wir schließen unsere Betrachtungen zur Anwendungsphase im Projekt k-med mit der Bemerkung, dass mit Hilfe der obigen http-Anfragen auch jedes WWW-basierte Lernsystem mit einem ontologiebasierten Glossar verbunden werden kann.

3.7 Prozessoptimierung

Der letzte Abschnitt dieses Kapitels widmet sich der Frage, wie der kooperative Ontologieerstellungsprozess nach Holsapple und Joshi verbessert und für Anwendungen, die ähnlich wie k-med auf die entstehende Ontologie als strukturiertes Vokabular einer Verschlagwortung zurückgreifen, optimiert werden kann. Für die vorliegende Arbeit wird vor allem die ähnlichkeitsbasierte Vorgehensweise des Ontologieanreicherungsverfahrens als entscheidend herausgestellt.
Die Vorgehensweise, die wir definiert und angewandt haben, als die Ontologie für das Projekt k-med erstellt wurde, war Gegenstand einer Umfrage unter beteiligten Medizinern. Dabei trat zutage, dass zwar die Funktionalität des eingesetzten Werkzeuges als hilfreich empfunden wurde, jedoch der Nutzen der Ontologie für die Projektbeteiligten nicht vollständig ersichtlich wurde. Dies war der Fall, obwohl in zahlreichen Schulungen und in der vorliegenden Demonstration der WWW-Ansicht der Ontologie als Glossar frühzeitig Elemente der Anwendungsphase für die Autoren bereitgestellt wurden.
Aus diesem Zusammenhang heraus deuten wir die Ergebnisse der Untersuchung wie folgt. Alle beteiligten Fachautoren erstellten zum ersten Mal eine Ontologie über ihr Fachgebiet, so dass der Nutzen nicht von vornherein offensichtlich sein kann, zumal erst parallel zur Ontologieentwicklung das eigentliche Unterrichtsmaterial in Form von Lernobjekten und Kursen hergestellt oder verfügbar wurde. Gleichzeitig erscheint es wichtig, initiale Ontologien mit einer rudimentären Ober- und Unterbegriffsstruktur zur Verfügung stellen zu können, da sonst eine geringere Beteiligung der Fachautoren in einem derartigen Projekt zu befürchten ist. Hiermit entstehen Fragen, die den weiteren Hergang der vorliegenden Arbeit entscheiden. Zum einen muss für eine Optimierung des Erstellungsprozesses nach Holsapple und Joshi ein automatisches Verfahren gefunden werden, das bereits in den Phasen der Verankerung und der iterativen Verbesserung zur möglichen Erweiterung bestehender Ontologien beiträgt, da eine reichhaltige, durch externe Standards wie die Gegenstandskataloge herstellbare Ober- und Unterbegriffsstruktur nicht immer gegeben sein muss, sondern vielmehr kleine oder sehr kleine Ausgangsstücke einer fachgebietsspezifischen Ontologie. Zum anderen sollte das Verfahren das Wechselspiel von Ontologieanwendung und iterativer Korrektur unterstützen. Hierbei ist insbesondere zu beachten, dass das mit Schlagworten zu versehende Material der Anwender einen Bezug zur Ontologie erhalten muss.
Technisch gesehen muss ein solches Verfahren für die Fachautoren transpa-

rent und ohne zusätzlichen Lernaufwand bezüglich vertiefender Arbeitsweisen der Wissensrepräsentation ablaufen. Das Verfahren sollte idealerweise noch Potiential für die kooperativen Aspekte der Erstellung in sich bergen, wie beispielsweise die vorausschauende Identifikation von fachlichen Überschneidungen oder sogar die Vermeidung von Konfliktfällen durch die gemeinsame Nutzung von Begriffsvorschlägen. Für den Moderator sollte ebenfalls keine zusätzlicher Einarbeitungszeit (zur Erfassung fachlicher Zusammenhänge) entstehen.
Wir folgern aus diesen Überlegungen, dass wir ein Verfahren zur automatischen Unterstützung des Erstellungsprozesses suchen, das aus elektronisch verfügbarem sprachlichem Material Hilfen bei der Identifikation relevanter Fachbegriffe liefert und diese Fachbegriffe in Form von Vorschlägen in eine vorhandene Ontologie, die immer den Gegenstand der Bearbeitung im kooperativen Prozess darstellt, einordnet. Zudem kann eine Erleichterung dadurch gefunden werden, dass Schlagworte der Autoren, die noch nicht in der Ontologie vorhanden sind, durch den gleichen Mechanismus als Begriffsvorschläge an der sinnvollsten Stelle in der Ontologie zugewiesen werden. Schließlich sollte für die Fachgebietsautoren eine der Semantik der Ontologie angepasste Möglichkeit bestehen, die Anzahl der Begriffsvorschläge zu regulieren.
Die Begriffsvorschläge sollten durch das automatische Verfahren in einer Weise gefunden werden, die mehrere Aspekte der kooperativen Ontologieerstellung zugleich erfasst:

- wir müssen ein Kriterium definieren, das sicher stellt, welche neuen Begriffe der potentiell sehr umfangreichen Menge an denkbaren Begriffen zur Erweiterung in Frage kommen

- ein neuer Begriff sollte an die Stelle der bestehenden Ontologie vorgeschlagen werden, die die stärksten semantischen Bezüge zum neuen Begriff aufweist

- gleichzeitig sind grundsätzlich mehrere solcher Stellen denkbar. Im Sinne der Kooperation ist es von besonderem Interesse, ob diese möglicherweise mehrfach vorhandenen Begriffsvorschläge in den Zuständigkeitsbereich verschiedener Autoren fallen.

- das Verfahren sollte Auskunft darüber geben können, ob bei mehreren Schlagwörtern, die noch nicht als Begriffe angelegt wurden, ein zu präferierendes hinsichtlich der Integration in die bestehende Ontologie identifiziert werden kann

Eine letzte wichtige Anforderung an die automatische Unterstützung von Ontologieerstellungsprozessen resultiert bereits aus der Definition 4, die die Gegebenheiten, die zur Verarbeitung durch ein automatisches Verfahren vorhanden sind, genau spezifiziert. Darüber hinaus sollen möglichst wenige unterschiedliche zusätzliche elektronische Eingangsmaterialien für das automatische Verfahren notwendig sein. Ist die der Fall, so erklärt sich eine automatische Erweiterung einer Ontologie in erster Linie aus ihrer Struktur.
Betrachten wir die Anforderungen des vorliegenden Abschnittes als Ganzes, so können wir sie auf der Grundlage von Definition 4 und ihrer beiden Zusatzannahmen durch **Einführung einer quantitativen oder qualitativen Ähnlichkeit zwischen Paaren von Begriffen** erfüllen. Diese Ähnlichkeitsfunktion sollte sowohl zwischen Begriffen innerhalb einer gegebenen Ontologie als auch zwischen einem externen, noch nicht in die Ontologie aufgenommenen Begriff (beziehungsweise einem potentiellen Begriffsbezeichner) definierbar sein. Die Ähnlichkeit ermöglicht es, für ein Wort seine besten Zuordnungen in die bestehende Ontologie zu finden. Ein Verständnis der Ähnlichkeit hilft ebenso dabei zu ermitteln, ob ein ausreichender semantischer Bezug zwischen dem Wort und einem oder mehreren Begriffen aus der Ontologie vorliegt und die Integration eines neuen Begriffes als sinnvoll zu erachten ist.
Der ausstehende Teil der vorliegenden Arbeit wird sich Lösungsvorschlägen, sowie der Spezifikation und Bewertung einer den Anforderungen genügenden ähnlichkeitsbasierten Methode widmen. Wir beachten dabei die Voraussetzungen, die durch Definition 4 und die im vorliegenden Abschnitt als Anforderungsliste identifizierte Verbesserungsmöglichkeit des kooperativen Erstellungsprozesses entstehen.

3.8 Bemerkung: eigene Beiträge

Die Erstellung einer Ontologie für das Projekt k-med wurde durch den Verfasser der vorliegenden Dissertationsschrift durchgeführt. Dabei ist die detaillierte Betrachtung eines durch vorstrukturiertes Ausgangsmaterial verankerten kooperativen Ontologieerstellungsprozesses in der vorgestellten Form neuartig.
Eine starke Formalisierung eines kooperativen Verfahrens zur Ontologieerstellung, wie wir sie durch die Ablaufdiagramme in der k-med-Fallstudie vorweisen können, bildet ebenfalls eine neuartige Erweiterung der Arbeiten von Holsapple und Joshi. Das Ziel unserer weiteren Argumentation wird ebenfalls ein neuartiger und originärer eigener Beitrag sein: wir haben An-

forderungen an eine automatische Unterstützung des Verfahrens aufgestellt. Entlang dieser Anforderungen sollen im Fortgang der vorliegenden Dissertationsschrift ein neuartiger Algorithmus zur Anreicherung bestehender Ontologien entwickelt und bewertende Untersuchungen desselben durchgeführt werden.

Kapitel 4

Verwandte Arbeiten

Das folgende Kapitel gibt einen Überblick über verwandte Arbeiten, in denen mit automatischen Verfahren Ontologien erstellt oder bestehende Ontologien erweitert werden. Für die automatischen Verfahren, die eine Ontologie vollständig generieren, werden wir die Frage untersuchen, ob die jeweilige Methode auf die im vorigen Kapitel geforderte automatische Erweiterung einer bestehenden Ontologie übertragbar ist.
Wir nehmen bei der Betrachtung verwandter Arbeiten folgende Eingrenzung vor. Beim Gang unserer Untersuchungen werden wir uns auf solche Arbeiten beschränken, die automatisch Begriffe und Ober- und Unterbegriffsbeziehungen herstellen, die somit mit der Menge B und der Halbordnung $\leq$ gemäß Definition 4 operieren [1] .
Wenn wir uns die Anforderungen, die wir als Schlussfolgerungen des letzten Kapitels herausgearbeitet haben, vergegenwärtigen, so sollte ein Verfahren zur automatischen Unterstützung des Erstellungsprozesses mittels einer Herstellung oder Erweiterung von B und $\leq$ mit bereits existierenden Ontologien als Ausgangsmaterial arbeiten können. Dies liegt bereits darin begründet, dass die initiale Ontologie oftmals aus vorhandenen Quellen fachlichen Wissens gewonnen werden kann oder die Integration dieser Quellen sogar eine Grundanforderung der Ontologieerstellung ausmacht. Gleichzeitig benötigt

[1] Definition 4 lässt als Anwendungsgebiet einer automatischen Ontologieerstellung auch noch die automatische Erzeugung von Relationen zur Herstellung oder Erweiterung der Relationenmenge R und ihrer Ausprägung σ nennen. Bei einigen der im Folgenden zu zeigenden Ansätze geschieht dies auch, allerdings unterscheiden sich die Techniken so stark, dass Verfahren zur automatischen Herstellung oder Erweiterung von R ein eigenes Forschungsgebiet darstellen (vergleiche hierzu [14]). Für die Durchführung eines prägnanten Vergleiches erscheint daher die Beschränkung auf Begriffe und ihre Halbordnung $\leq$ notwendig.

ein Verfahren, das mit B und der Unterbegriffsrelation $\leq$ operiert, zusätzliche Informationen als Eingangsparameter. Alle der im folgenden Kapitel zu untersuchenden Verfahren werden zusätzliche Informationen aus so genannten Textkorpora beziehen. Wir grenzen uns damit von Verfahren, die Wissensrepräsentationen aus Datenbanken lernen (wie zum Beispiel [30]), ab. Wir definieren einen Textkorpus nach [91] wie folgt:

Definition 6 (Textkorpus) *Ein Textkorpus ξ ist eine Sammlung von natürlich vorkommenden natürlichsprachlichen Texten, die ausgewählt wurden um den Zustand oder die Reichhaltigkeit einer Sprache aufzuzeigen oder zu untersuchen.*

Weitergehende Definitionen schließen auch Sammlungen gesprochener Sprache mit ein, hiervon wollen wir jedoch absehen oder zumindest für gesprochene Korpora die Existenz einer Transkription in Textform annehmen. Die Definition umfasst weder Sammlungen formaler Sprache noch im engeren Sinne rechnergenerierte natürliche Sprache. Letztere wird beispielsweise bei rechnergestützten Spielen alternativ zu Grafiken als Zusammenfassung oder Darstellung von Spielverläufen eingesetzt. Ein Beispiel findet sich unter [39]. Für die Fragestellungen der vorliegenden Arbeit treffen wir folgende Zusatzannahme.

Annahme 3 (Existenz der Bezeichner im Textkorpus) *Der Textkorpus ξ ist für die automatische Konstruktion oder Erweiterung einer Ontologie sprachlich relevant. Insbesondere enthält ξ im Falle der automatischen Erweiterung einer bestehenden Ontologie deren Begriffsbezeichner.*

Textkorpora nach Definition 6 und Annahme 3 bilden eine andere Grundlage für automatische Ontologieerstellungsverfahren als bereits vorstrukturiertes Material, wie es in den Ansätzen von [67] (die XML als Zusatzinformation ausnutzen) oder [20] (wo RDF verwendet wird) Eingang findet.

4.1 Theoretische Ansätze

Den in den Abschnitten 3.1.1 und 3.1.2 dargestellten theoretischen Ansätzen ist gemein, das sie ein statistisches Sprachmodell [59] der Wörter in einem Textkorpus zugrunde legen. Ein statistisches Sprachmodell legt initial eine Anzahl von Merkmalen (Attribute) fest; die empirischen Häufigkeiten dieser Merkmale werden sodann im Textkorpus für einzelne Wörter überprüft und

meistens als Vektor dargestellt.
Ansätze mit einem statistischen Sprachmodell unterscheiden sich somit prinzipiell von so genannten symbolischen Ansätzen, die im weiteren Sinne Regeln für die Auswertung natürlichsprachlicher Sätze aus einem Textkorpus aufstellen und aus diesen Ontologien generieren. Mit anderen Worten handelt es sich bei symbolischen Ansätzen um am natürlichsprachlichen Satz orientierte, meist auf direkte Umsetzung seiner Subjekt-Prädikat-Objektstrukturen fokussierte Methoden. Sehr einfache Beispiele für symbolische Ansätze wären die Umsetzung der natürlichsprachlichen Sätze

'Franz ist Konditor.'

und

'Konditoren brauchen Zucker.'

in Relationen der Art $Franz \leq Konditor$ und $brauchen(Konditor, Zucker)$.
Bei der Darstellung von TextStorm und Clouds zu Ende des Kapitels werden wir anhand eines weiteren Beispiels einen solchen symbolischen Ansatz und seine Problematik zeigen.
Für das statistische Sprachmodell nehmen wir im Folgenden gemäß der Schreibweise in [59] an, dass es eine Menge von Attributen $\mathcal{A} = (\mathcal{A}_1, ..., \mathcal{A}_N)$ gibt. Als Beispiel für ein auf einen Textkorpus ξ bezogenes Attribut ist das Vorkommen mit einem bestimmten Wort W in einem Satz in ξ zu nennen. Wenn wir N solcher Wörter W_1,...W_N vorgäben, so erhielten wir ein statistisches Sprachmodell $\mathcal{A} = (\mathcal{A}_1, ..., \mathcal{A}_N)$.
Ferner sei für jedes Wort oder jede festgelegte Wortgruppe x aus einem Textkorpus ξ durch einen Vektor $\vec{x} = (x_1, ..., x_N)$ mit natürlichen Zahlen $x_1, ..., x_N$ angegeben, wie oft x die Attribute erfüllt: x_i zeigt dabei an, wie oft das Attribut $\mathcal{A}_i$ erfüllt ist. Diese Gegebenheiten sind für die automatischen Konstruktionsansätze von Ontologien im Folgenden gleich. Wie wenden uns nun vier Gruppen von theoretischen Ansätzen zu, nämlich der Clusterbildung als nicht überwachten Ansatz, dem Kategorisieren als überwachtem Ansatz und einer möglichen Übertragung einer Technik der formalen Begriffsanalyse auf automatische Ontologiekonstruktionen. Schließlich werden wir auch eine rein auf die Kollokationen, im Textkorpus bezogene Methode erläutern, die eine gegebene Ontologie voraussetzt.

4.1.1 Hierarchische Clusterbildung

Clusterbildung ist ein Forschungsgebiet, das auf verschiedenste Gegenstände angewandt wird [57]. Es stellt eine nicht überwachte Methode des maschi-

nellen Lernens in dem Sinne dar, dass vor der Anwendung eines Clusteralgorithmus keine Anhaltspunkte über die Zusammengehörigkeit oder Verwandtschaft bestimmter Attributstrukturen oder der Clusterbildung unterzogenen Gegenstände vorliegen [71].
Analog zu unserer Annahme eines statistischen Sprachmodells seien für die aus Wörtern bestehende Gegenstandsmenge G wieder Attribute $\mathcal{A}$ und für jedes $g \in G$ die numerischen Angaben über die Erfüllung dieser Attribute als Vektor $\vec{g}$ gegeben. Ziel herkömmlicher Clusterverfahren ist die Zusammenfassung von Gegenständen zu einem so genannten Cluster. Jedem Gegenstand wird dabei ein Cluster zugeordnet. Man benötigt dafür eine Ähnlichkeitsfunktion s_0 auf den Paaren der Gegenstandsmenge G:

$$s_0 : G \times G \mapsto \mathbb{R}_0^+. \tag{4.1}$$

Die Gruppierung von Gegenständen, die untereinander einen hohen Ähnlichkeitswert aufweisen, nennt man Clusterbildung. Sowohl bei der Definition der Ähnlichkeitsfunktion mittels der Vektoren $\vec{g}$ als auch bei der eigentlichen Bildung der Gruppen existieren mannigfaltige Ansätze. So wird bei der Ähnlichkeitsfunktion oftmals auf das euklidische Winkelmaß zurückgegriffen werden, wobei kleine Winkel mit einer hohen Ähnlichkeit einhergehen. Eine ausführliche Darstellung alternativer Ähnlichkeitsmaße - insbesondere für Sprachdaten - folgt in Kapitel 5. Die Alternativen bei der Bildung von Clustern bestehen hauptsächlich in der Abfolge der Vergleiche durch s_0. Die unterschiedliche Wahl von Repräsentanten $\vec{c}$ bereits vorhandener Cluster c bei der sukzessiven Zuordnung neuer Gegenstände (bei Ausdehnung oder Unterteilung der Gegenstandsmenge während des Clustervorgangs) bildet die zweite prinzipielle Wahlmöglichkeit beim Design eines Clusteralgorithmus. Wenn wir die Potenzmenge von G als $P(G)$ schreiben, so kann diese Wahlmöglichkeit als Definition einer weiteren Ähnlichkeitsfunktion

$$s_1 : \{\vec{c}|c \in P(G)\} \times G \mapsto \mathbb{R}_0^+. \tag{4.2}$$

betrachtet werden, die in der Regel auf s_0 beruht.
Für die automatische Konstruktion von Ontologien werden Formen des so genannten hierarchischen Clusterns angewandt. Die Gegenstände in $G := \{g_1, ..., g_n\}$ sind in diesem Falle Wörter oder Wortgruppen mit der zugehörigen Vektordarstellung $\vec{g_i}$ bezüglich einer Menge von Attributen $\mathcal{A} = (\mathcal{A}_1, ..., \mathcal{A}_N)$ und einem Textkorpus ξ. Des Weiteren benötigen wir eine weitere Ausdehnung der Ähnlichkeitsfunktionen s_0 und s_1 zu

$$s_2 : \{\vec{c}|c \in P(G)\} \times \{\vec{c}|c \in P(G)\} \mapsto \mathbb{R}_0^+. \tag{4.3}$$

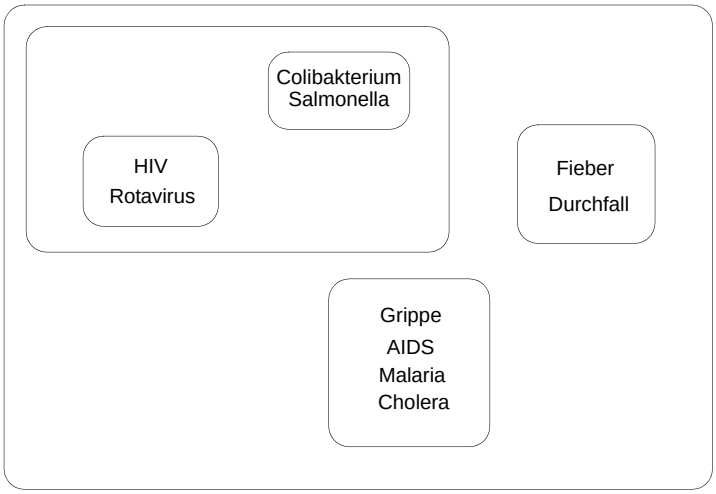

Abbildung 4.1: Hierarchisches Clustern

Durch s_2 wird es möglich, das Ergebnis des ersten Clusteralgorithmus wiederum als Gegenstandsmenge einer Clusterbildung zu verstehen. Somit können dann Cluster zu neuen Clustern zusammengefasst werden[1]. Der Vorgang ist iterativ anwendbar, bis nur noch ein mit der Gegenstandsmenge (den Wörtern oder Wortgruppen aus) G identisches Cluster übrig ist. Hierbei ist zu beachten, dass das Verfahren so definiert sein muss, dass dieses letzte mit G identische Cluster per Definition des Verfahrens auch in endlicher Zeit hergestellt werden kann. Das heißt insbesondere, dass dies in endlich vielen Iterationsschritten erfolgen muss. Abbildung 4.1 zeigt exemplarisch das Resultat des hierarchischen Clusterns und den allgemeinen Zusammenhang zur automatischen Ontologiekonstruktion [11]. Gegeben ist hier die Gegenstandsmenge

$$G_0 := \{Colibakterium, Salmonalla, HIV, Rotavirus,$$

[1]Unsere Darstellung zeigt den zusammenfassenden Clusteransatz, das so genannte bottom-up Verfahren. Clustern kann auch durch Unterteilung erreicht werden. Dies bezeichnet man als top-down Ansatz, vergleiche[68].

$$Grippe, AIDS, Fieber, Durchfall\}.$$

Eine hohe Ähnlichkeit

$$s_0(Colibakterium, Salmonella) \tag{4.4}$$

mit einer für den vorliegenden Clustervorgang definierten Ähnlichkeitsfunktion s_0 führt dazu, dass diese beiden Wörter in einem Cluster zusammengefasst werden. Dieses Cluster ist links oben in der Abbildung 4.1 dargestellt. Die Tatsache, dass aus $G\backslash\{Colibakterium, Salmonella\}$ kein Wort mehr zu finden ist, dessen Vektorrepräsentation eine hinreichend hohe Ähnlichkeit s_1 zum Cluster $\{Colibakterium, Salmonella\} := C_{01}$ aufweisen konnte, führte in der ersten Iterationsstufe des hierarchischen Clusterns kein weiteres Wort in dieses Cluster. In der ersten Iterationsstufe des hierarchischen Clusterns entstehen auf ähnliche Weise die Cluster $\{HIV, Rotavirus) := C_{02}$ und $\{Fieber, Durchfall) := C_{03}$. Beim Cluster

$$\{Grippe, Aids, Malaria, Cholera) := C_{04}$$

gab es eine Zwischenstufe, bei der mit Hilfe von s_1 ein ursprünglich vorhandenes Cluster $\{x, y\}$ mit $x, y \in \{Grippe, Aids, Malaria, Cholera\}$ erweitert wurde, da s_1 für die beiden außer x und y verbleibenden Wörter einen hinreichend hohen Wert aufwies.
In unserem Beispiel kommt es dann in der nächsten und allen weiteren Iterationsstufen zur Anwendung der Ähnlichkeitsfunktion s_2 auf

$$\{C_{01}, C_{02}, C_{03}, C_{04}\} \times \{C_{01}, C_{02}, C_{03}, C_{04}\}.$$

Lediglich die Ähnlichkeit zwischen C_{01} und C_{02} wurde als hoch genug befunden um diese wieder zu einem neuen Cluster C_{11} zusammenzufassen. Für

Tabelle 4.1: Cluster und Begriffe

Cluster	*Begriffsbezeichner*
C_{01}	Bakterium
C_{02}	Virus
C_{03}	Krankheit
C_{04}	Symptom
C_{11}	Krankheitserreger
G_0	$\top$

den Rest des Iterationsdurchgänge wurde kein Ähnlichkeitswert von s_2 mehr

gefunden, der hoch genug gewesen wäre, um zwei oder drei der bislang entstandenen Cluster zu größeren Clustern zusammenzufassen. Somit endet das Verfahren, indem G_0 als alles zusammenfassendes Cluster angegeben wird.
Wenn wir nun die Wörter aus der Gegenstandsmenge und die einzelnen Cluster mit Begriffen identifizieren, so können wir, wie beispielsweise in den Arbeiten von [13] gezeigt, zur Struktur einer Ontologie gelangen, die Definition 4 genügt. Lediglich die natürlichsprachlichen Bezeichner der Begriffe sind noch anzugeben. Im Einzelnen fassen wir die Wörter aus G_0 als Bezeichner der speziellsten Begriffe innerhalb der Ontologie auf, das heißt als Bezeichner der Begriffe, die keinen weiteren Unterbegriff mehr besitzen. Darüber hinaus ordnen wir die in Tabelle 4.1 aufgeführten Bezeichner zu.
Die Ober- und Unterbegriffsrelation $\leq$ wird über Mengenzugehörigkeiten erklärt. Die Transitivität ergibt sich aus der Tatsache, dass Cluster späterer Iterationsstufen die Cluster früherer Iterationsstufen beinhalten. Somit gilt beispielsweise per Konstruktion

$$\{Salmonella\} \subseteq C_{01} \subseteq C_{11} \subseteq G_0, \tag{4.5}$$

was sich direkt in Unterbegriffsbeziehungen der Form

$$Salmonella \leq Bakterium \leq Krankheitserreger \leq \top \tag{4.6}$$

übertragen lässt. Die Zuordnung von Clustern zu Begriffen (mit natürlichsprachlichen Bezeichnern), die eingeführte Unterbegriffsrelation $\leq$ und die Einführung des Oberbegriffes $\top$ durch die Identifikation von G_0 (und im Allgemeinen der gesamten Gegenstandsmenge G) mit $\top$ stellen somit aus dem Ergebnis hierarchischer Clusterverfahren Ontologien im Sinne der Definition 4 her.
Arbeiten, die auf solchen Clusterverfahren beruhen, finden sich vor allem in der Konstruktion automatischer Thesauri (siehe [86], [19]). Dort werden aus allgemeinen Textkorpora hierarchisch gegliederte Cluster erzeugt, die den Sinnkategorien eines Thesaurus entsprechen. Auch Lagus Arbeiten zur Clusterbildung finnischer Verben [58] durch Kohonenkarten [57] zeigen, dass solche Verfahren vor allem bei allgemeineren Gegenstandsmengen (in diesen Fällen: Wortmengen) als Erfolg versprechend anzusehen sind. Lagus betrachtet allerdings nur eine Ebene des Clusterns und ermittelt, ob Verben ähnlicher Tätigkeit (beispielsweise Verben zum Thema Fortbewegung wie 'laufen, 'fliegen' et cetera) in ein Cluster abgebildet werden. Bisson, Nedellec und Cañamero haben in ihrer Arbeit zur korpusbasierten Ontologieerstellungsumgebung darauf hingewiesen [10], dass die Übertragung dieser Prinzipien auf fachgebietsspezifische Ontologien und auf speziellere Textkorpora eine Herausforderung darstellt, da die Korpusgrößen im Gegensatz zu

allgemeinen Textkorpora kleiner ausfallen können und dieser Aspekt noch wenig erprobt wurde.
Die im folgenden Abschnitt darzustellenden Verfahren benutzen im Gegensatz zu den nicht überwachten Clusterverfahren zusätzliche Eingangsinformationen.

4.1.2 Überwachte Verfahren

Im Gegensatz zu den im vorherigen Abschnitt vorgestellten Clusterverfahren gehört es zu den Grundprinzipien überwachter Verfahren, dass für eine bereits vorhandene Menge an Daten semantische Interpretationen vorhanden sind. Es existiert somit für eine so genannte Trainingsmenge bereits das erwünschte und richtige Vorhersageergebnis [71]. Oftmals wird dies bei der Kategorisierung von vektorwertigen Objekten benutzt.
Übertragen auf das Kategorisieren von Wörtern nach auf einen Textkorpus bezogenen Attributen bedeutet dies Folgendes. Es sind die Attributmenge $\mathcal{A} = (\mathcal{A}_1, ..., \mathcal{A}_N)$ und die für jedes Wort und jede Wortgruppe x aus dem Textkorpus ξ durch einen Vektor $\vec{x} = (x_1, ..., x_N)$ quantitativ erfassten Eigenschaften gegeben. Außerdem existieren auch noch für eine Menge von Wörtern W_t zusätzliche Zuordnungen z_0 zu Elementen einer weiteren Menge K:

$$z_0 : W_t \mapsto K. \tag{4.7}$$

Man nennt W_t die Trainingsmenge, K die Kategorien und z_0 die Zuordnung zu den Kategorien. Kategorien können beispielsweise Oberbegriffe des jeweiligen Wortes sein, was bereits den Hinweis auf den Zusammenhang zu automatischen Verfahren der Ontologiekonstruktion liefert. Typischere Anwendungen, die sich etwa bei [16] finden, ist die Klassifikation von durch Vektoren repräsentierte Textdokumente. Analog zu den Attributen $\mathcal{A}$ (für unsere Probleme aus dem Bereich der Auswertung von Eigenschaften eines Wortes im Textkorpus) müssen für die Vektorrepräsentation des Dokumentes feste Attribute existieren. Es handelt sich bei besagten Anwendungen auf Dokumentenbasis etwa um die Klassifikation von Webseiten nach den Kategorien der Suchmaschinen Google[32]. Auch das Sortieren von Spam-E-Mails kann als Kategorisierungsproblem aufgefasst werden. Die Zuordnung z_0 ist für die Trainingsmenge stets eine surjektive Abbildung. Kategorisierungsverfahren leiten auf systematische Weise aus der Zuordnung der Trainingsmenge eine Ausdehnung z_1 der Zuordnung auf beliebige, anhand der Attribute $\mathcal{A}$ repräsentierte Objekte ab.
Für die Entstehung der Trainingsmenge ist festzuhalten, dass sie entweder aus explizit vorgenommenen Zuordnungen von Anwendern stammt oder aus

ex-post-Betrachtungen abgeleitet wird. Explizit vorgenommene Zuordnungen lägen etwa dann vor, wenn eine (möglicherweise offene) Gruppe von Nutzern fortwährend Webseiten thematischen Kategorien zuordnet und regelmäßige Aktualisierungen der Zuordnung in die automatische Kategorisierung einflössen. Ex-post-Betrachtungen könnten sich beispielsweise darauf beziehen, wie schnell ein Benutzer eine eingehende e-Mail löscht, was in diesem Falle durch eine vom Benutzer akzeptierte Anwendung zur zeitlichen Erfassung des Löschvorgangs hinausliefe.
Die Nutzung der überwachten Klassifikationsverfahren für eine automatische Konstruktion von Ontologien ist vergleichsweise neu. Hofmann geht in aktuelleren Arbeiten [44] darauf ein, wie hierarchische Beziehungen der Kategorien oder auch Informationen über die semantische Ähnlichkeit von Kategorien in Informationen für das Klassifizieren durch Supportvektormaschinen nach [50] und [18] überführt werden können. Die Arbeiten von Hofmann gehen dabei stets davon aus, dass bereits eine Ontologie im Sinne der Minimalanforderungen nach Definition 4 existiert. Dabei wird dann bei einer Zuordnung neuer Wörter oder Wortgruppen auf die Kombination zweier Techniken abgestellt. Zum einen schlägt sich der Fall, dass ein Wort zu einer Kategorie (in der überwachten Interpretation: einem Begriff nach Definition 4) gehört, für die Trainingsmenge und für die Menge der zu klassifizierenden Wörter auch in den Attributen nieder, die die Oberkategorien (Oberbegriffe) des Wortes bilden. Zum anderen wird für die Trainingsmenge eine Fehlklassifikation durch z_1 dann als weniger störend bewertet, wenn ein Wort zwar nicht richtig, aber zumindest zu einem Oberbegriff der eigentlich richtigen Kategorie zugeordnet wurde.
Hofmann hat zu den mit Hilfe dieser überwachten Strategien formulierten Verfahren Experimente durchgeführt, die für vorhandene Thesauri Neuklassifikationen einzelner Wörter vornahmen. Diese Struktur der Trainingsmenge und des gesamten experimentellen Aufbaus von Hofmann werden wir in unserer Besprechung der theoretischen Ansätze kritisch betrachten.
Zunächst jedoch stellen wir im folgenden Abschnitt rein kollokationsbasierte Verfahren vor, die hauptsächlich das direkte gemeinsame Vorkommen zweier Wörter in einem Textkorpus auswerten.

4.1.3 Direkte Kollokationsauswertung

Im Unterschied zu den auf hierarchischer Clusterbildung oder auf überwachter Klassifikation beruhenden Verfahren werden bei Verfahren mit einer direkten Kollokationsauswertung keine fest begrenzten Attributmengen $\mathcal{A}$ und folglich auch keine Vektordarstellungen zuzuordnender oder zu gruppieren-

der Begriffe verwendet. Auch die Zuordnungs- und Ähnlichkeitsfunktionen sind nicht wie bei den beiden bislang vorgestellten theoretischen Ansätzen definiert. Stattdessen untersuchen diese Verfahren für eine gegebene Ontologie Ω die Wörter, welche gemeinsam mit den natürlichsprachlichen Bezeichnern der Begriffe aus Ω in einem Textkorpus auftreten. Wir definieren diese Wörter als so genannte Kollokatoren und die Tatsache des gemeinsamen Auftretens als Kollokation. Die unterschiedlichen in der computerlinguistischen Literatur auftretenden Redeweisen über Kollokatoren schränken wir dabei zu Gunsten eines statistischen Ansatzes ein.

Definition 7 (Kollokator, Kollokation) *Sei ξ ein Textkorpus und seien x und y Wörter, die in ξ auftreten. Sei außerdem ein festes Kriterium vorgegeben, anhand dessen entschieden werden kann, ob x und y gemeinsam auftreten. Ist dieses Kriterium erfüllt, so unterliegen x und y einer Kollokation. Wir nennen dann x einen Kollokator von y und y einen Kollokator von x.*

Die in der Definition angesprochenen festen Kriterien können daraus bestehen, dass Wörter x und y im gleichen Satz, im gleichen Abschnitt, im gleichen Textdokument oder innerhalb eines genau definierten Abstandes (mit einer festgelegten Höchstanzahl von Wörtern zwischen x und y) auftreten.
Es folgt ein Beispiel für die direkte Verwertung von Kollokationen zur Erweiterung von Ontologien. Seien *Krankheit* und *Durchfall* Begriffe einer Ontologie. Im Beispiel betrachten wir die folgenden Kollokationen zu den natürlichsprachlichen Bezeichnern aus einem Textkorpus ξ_{ex}, wobei Konjunktionen, Präpositionen, Pronomen und Hilfsverben keine Berücksichtigung fanden.

Krankheit:
Herz (1543), AIDS (1063), Patienten (962), Alzheimer (773), Krebs (544), koronar (429), Blut (354), Parkinson (311), Symptome (304), Virus (300)
Durchfall:
Symptome (30), Virus (11), Erbrechen (5), Übelkeit(3), Salmonella(1)

Die Beispieldaten sind folgendermaßen zu interpretieren: 'Herz' trat 1543 im gleichen Satz mit dem Kollokator 'Krankheit' (also dem natürlichsprachlichen Bezeicner des Begriffes *Krankheit*) in ξ_{ex} auf, 'Symptome' mit 'Durchfall' (dem natürlichsprachlichen Bezeichner des Begriffes *Durchfall*) 30 Mal im gleichen Satz in ξ_{ex} auf und so fort.
In der Literatur sind im Zusammenhang mit automatischen Ontologiekonstruktionen verschiedene Verfahren zur Auswertung dieser Kollokationsin-

formationen bekannt. Yarovsky [109] und Agirre [1] betrachten jeweils prinzipiell die Kollokatoren, die den beiden Begriffen nicht gemein sind, ergo alle außer 'Symptome' und 'Virus'. Die verbliebenen Kollokatoren stellen für Yarovsky und Agirre mögliche Erweiterungen der Ontologie dar. Im Gegensatz zu unserem Beispiel sind dazu allerdings aufwändigere Berechnungen nötig, da beide Autoren pro Begriff von Synonymlisten als Begriffsbezeichner ausgehen. Dies ist legitim, da beide Arbeiten wiederum mit Thesauri als Ontologien arbeiten. Yarowsky [109] berechnet in diesem Zusammenhang Auftrittswahrscheinlichkeiten. Heyer [42] hingegen unterscheidet Wortarten (Substantive und Verben) und ermittelt die statistische Signifikanz der Kollokation, um dann auf verschiedene Arten semantischer Relationen zu schließen. Dabei werden die Kollokatoren der Begriffsbezeichner getrennt voneinander ausgewertet, was laut Heyer für Hauptwörter tendenziell Unterbegriffe liefern kann.

4.1.4 Formale Begriffsanalyse

Neben beschreibungslogischen Ansätzen [98], die im Zusammenhang mit der automatischen Ontologiekonstruktion allerdings nur für sehr spezielle Textkorpora in Form von Patientenkarteien angewandt wurden, und hypothesenbasierten Verfahren [40], die eine intensive grammatikalische Aufbereitung des Textkorpus erfordern, existieren Methoden aus dem Bereich der formalen Begriffsanalyse. Diese Konstruktionsmethoden basieren auf logischen Verfahren. Im Gegensatz zu den bisher dargestellten Verfahren spielt hier weder die Anzahl der Kollokationen, noch die Darstellung von Wörtern als Vektor eine Rolle.
Die formale Begriffsanalyse, die wir in Kapitel 2 bereits eingeführt haben, um die Transitivität der Unterbegriffsrelation zu beschreiben, wurde in aktuelleren Arbeiten in den Zusammenhang mit Textkorpusanalysen gebracht. Stumme [97] wendet die formale Begriffsanalyse darauf an, zwei bestehende Ontologien miteinander zu einer Gesamtontologie zu verschmelzen. Obwohl dieses Verfahren namens FCA-Merge (FCA steht für Formal Concept Analysis, der englischsprachigen Bezeichnung für formale Begriffsanalyse) insofern nicht direkt als automatisches Konstruktionsverfahren anzusehen ist, konstruiert Stumme dabei Begriffsvorschläge für die Ausgangsontologien. Dies weist hinreichende Verwandtschaft mit den Forderungen aus unserem Kapitel über Ontologieerstellungsprozesse auf, da zusätzliche Begriffe für bestehende Ontologien auftreten können. Daher werden wir hier kurz auf FCA-Merge eingehen.
Gegeben seien zwei Ontologien Ω_1 und Ω_2 im Sinne der Definition 4. Zudem

sei ein Textkorpus ξ gegeben, der aus einzelnen voneinander unterscheidbaren Textdokumenten besteht. Des Weiteren seien für $i = \{1, 2\}$ die zur jeweiligen Ontologie Ω_i gehörigen Begriffe B_1 Merkmale eines formalen Kontexts K_i. Die Gegenstände G_j von K_i sind ist die jeweils gleiche Menge von Textdokumenten, ein Merkmal ist erfüllt, wenn ein Begriff in einem Dokument auftritt. Stumme spricht hier von der Instanz eines Begriffes, was einen weiteren Vorverarbeitungsschritt notwendig macht, beispielsweise in Form einer (möglicherweise automatischen) Verschlagwortung der Textdokumente durch die Begriffe aus B_1 und B_2.
FCA-Merge stellt aus K_1 und K_2 einen Gesamtkontext K her. Begriffe mit gleich lautendem natürlichsprachlichen Bezeichner werden in der Merkmalsmenge M des Gesamtkontexts unterschieden, das heißt, wir nehmen ohne Beschränkung der Allgemeinheit an, dass

$$K_1 \cap K_2 = \emptyset, M = K_1 \cup K_2, \tag{4.8}$$

Der Gesamtkontext definiert sich dann als $K := (G, M, I)$ durch Injektion der jeweiligen Merkmalsrelationen aus K_1 und K_2.
Aus K ergeben sich nun **formale** Begriffe wie in den Definitionen 2 und 3 festgelegt. Die **formalen** Begriffe müssen an dieser Stelle ausdrücklich von den Begriffen, die durch die Ontologien Ω_1 und Ω_2 gegeben sind, unterschieden werden. Zu jedem so entstandenen formalen Begriff (A, B) ist nun nach [104] eine Schlüsselmenge $S \subseteq M$ identifizierbar, für die für alle $X \subseteq S$ mit den Ableitungsregeln aus Abschnitt 2.2.1 gilt:

$$(S', S'') = (A, B), (X', X'') \neq (A, B) \tag{4.9}$$

Stumme unterscheidet nun für die formalen Begriffe (A, B) anhand der Mächtigkeit der jeweils zugehörigen Schlüsselmengen S_j vier Fälle. Durch die Fallunterscheidung wird dann von den formalen Begriffen des Kontexts K auf den Umgang mit den Begriffen aus Ω_1 und Ω_2 geschlossen.
Zwei der besagten Fälle ($|S_j| = 1$ für (A, B)) führen zu einer Verschmelzung oder fortzuführenden Trennung der ursprünglich in Ω_1 und Ω_2 vorhandenen Begriffe, einer der Fälle führt zu einem Begriffs- oder Relationsvorschlag für diejenigen, die eine Verschmelzung der Ontologie vornehmen. Stumme legt nahe, dass ein Hintereinanderschreiben der Begriffsbezeichner aus den Schlüsselmengen als Begriffsbezeichner des Vorschlags angeboten wird. Im hier beispielhaft gewählten Falle der Schlüsselmenge $S = \{Bakterium, Krankheit\}$ würde

$$BakteriumKrankheit$$

angeboten. Alternativ besteht die Möglichkeit, zwischen *Bakterium* und *Krankheit* eine Relation anzulegen.

4.1.5 Symbolische Ansätze

Der folgende Abschnitt stellt eine praktische Umsetzung der textbasierten Ansätze aus dem vorausgegangenen theoretischen Teil vor. Die vorzustellenden Implementierungen tragen Züge eines hybriden Systems, das mehrere so genannte symbolische Ansätze mit statistischen Ansätzen kombiniert.
Der Systemverbund TextStorm und Clouds wurde im Rahmen des Dr. Divago Projekts entwickelt[80]. Das Projekt zielte auf die semi-automatische Konstruktion semantischer Netze aus fachspezifischen Texten ab. Es setzt sich aus zwei Modulen zusammen, TextStorm und Clouds. TextStorm ist ein Werkzeug zur Verarbeitung natürlicher Sprache, welches mit Hilfe syntaktischer Ansätze binäre Ausdrücke aus einem Text extrahiert.

Die Ausdrücke, die TextStorm schließlich liefert, sind allesamt Phrasen, die aus einem Verb samt Subjekt und direktem Objekt bestehen. Das Verb spezifiziert hierbei die Relation zwischen den beiden Begriffen, die durch Subjekt und Objekt dargestellt werden. In unserer Notation liefern Subjekte und Objekte somit insgesamt gesehen die Menge B und Verben die Menge R im 4-Tupel, das die Ontologie definiert. Abbildung 4.2 gibt einen Überblick über das TextStorm Werkzeug. TextStorm benötigt keine Informationen über das zu bearbeitende Wissensgebiet. Die gefundenen Ausdrücke dienen dem Modul Clouds als Eingangsparameter. Das Verhalten von TextStorm lässt sich am einfachsten anhand eines Beispiels zeigen. Angenommen die Text-Datei am Eingang von TextStorm enthält den Satz

'Kühe, genau wie Hasen, essen nur Pflanzen, während Menschen auch Fleisch essen.'

Dann sollte TextStorm die Ausdrücke

essen(Kuh,Pflanzen)

essen(Hase,Pflanzen)

essen(Mensch,Pflanzen)

essen(Mensch,Fleisch)

extrahieren. In den dazu nötigen Schritten sollen Synonyme erkannt und Zweideutigkeiten aufgelöst werden. Auch die Problematik von Pronomen,

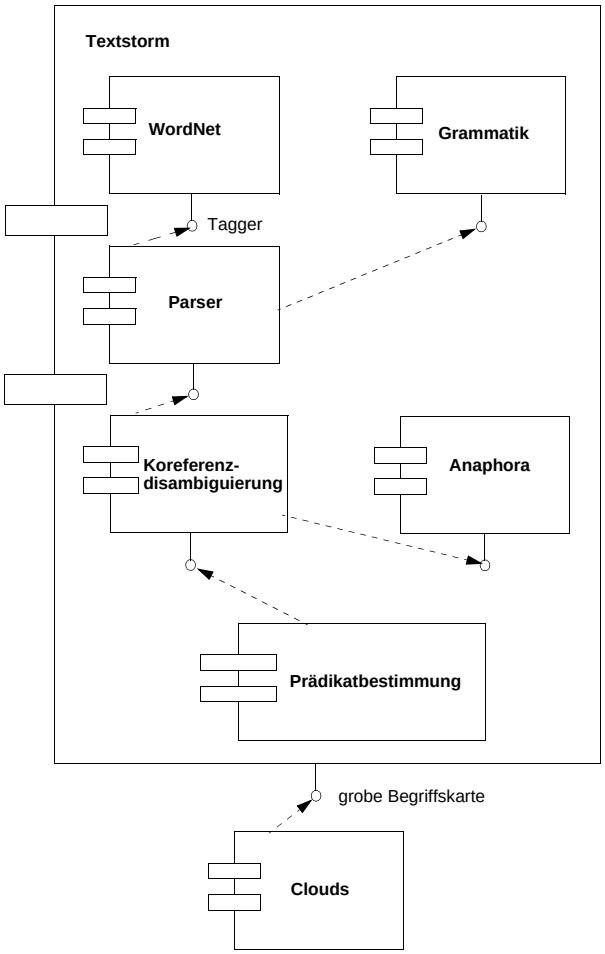

Abbildung 4.2: Übersicht TextStorm, Quelle: Perreira[80]

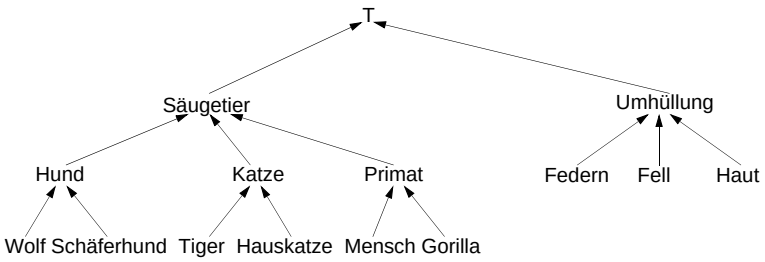

Abbildung 4.3: Halbordnung der Begriffe in Clouds, in Anlehnung an Perreira[80]

die sich auf Subjekte oder Objekte beziehen soll in diesen Schritten gelöst werden. Schließlich übergibt TextStorm Vorschläge für Begriffe und Relationen (Raw Conceptual Map) an Clouds.
In Interaktion mit dem Benutzer verfeinert Clouds die von TextStorm erhaltene Begriffskarte (Raw Conceptual Map). Eine Begriffskarte (Raw Conceptual Map) ist eine vereinfachte Form eines semantischen Netzes und besteht aus binären Ausdrücken, die Relationen zwischen zwei Begriffen darstellen. Basierend auf den Relationen und Begriffen, die der Benutzer eingibt, erzeugt Clouds schrittweise die korrespondierende Hierarchie. Die von TextStorm geleistete Vorarbeit erleichtert diesen Prozess und verringert die Informationen, die vom Benutzer an Clouds übergeben werden müssen.

Während der Arbeit mit Clouds werden dem Benutzer Fragen zu vermuteten Relationen oder neuen Begriffen gestellt. Die Abbildung 4.3 zeigt die Ableitung neuer Unterbegriffsrelationen durch Clouds. Die Arbeit durch den Benutzer erfolgt deduktiv. Angenommen die Relationen

$$hat(Tiger, Fell) \tag{4.10}$$

$$hat(Hauskatze, Fell) \tag{4.11}$$

$$hat(Hund, Fell) \tag{4.12}$$

seien durch TextStorm bereits aus dem Textkorpus extrahiert worden. Aus den von TextStorm erhaltenen binären Ausdrücken können von Clouds neue Relationen abgeleitet werden. Aus den positiven Beispielen 4.10, 4.11 und 4.12 werden solange Hypothesen gewonnen, bis ein negatives Beispiel wie etwa

$$hatnicht(Mensch, Fell) \tag{4.13}$$

der Hypothese widerspricht. In unserem Fall heißt das, dass zunächst für alle Begriffe auf der untersten (speziellsten) Ebene eine Unterbegriffsrelation zu Säugetier hergestellt wird, wenn dies vorher für mindestens einen der Begriffe Tiger, Hund oder Hauskatze gegeben war. Für den Fall des Begriffes Wolf wird vom Benutzer umgekehrt erfragt, ob er die Relation

$$hat(Wolf, Fell) \tag{4.14}$$

übernehmen will. Insgesamt existieren somit für den Benutzer zwei Aktionen bezüglich der Begriffe. Diese wird dann geändert, um den neuen Bedingungen zu entsprechen. Das TextStorm- und Clouds-System soll das Bilden von Taxonomien vereinfachen, aber nicht automatisieren. Die Hauptverantwortung trägt immer noch der Benutzer.

4.2 Diskussion der Ansätze

Das Ziel der vorliegenden Arbeit besteht in einer ähnlichkeitsbasierten automatischen Unterstützung von kooperativen Ontologieerstellungsprozessen. Wie wir im Kapitel 3 dargelegt haben, soll das Verfahren, welches wir zur Verwirklichung der Zielvorstellung definieren wollen, bereits bestehende Ontologien verarbeiten und um neue Begriffe anreichern können. Die Kritik, die wir nun an den im Einzelnen dargestellten Verfahren äußern, ist somit vor dem Hintergrund dieses festgelegten Ziels zu verstehen.
Symbolische Vorgehensweisen, wie wir sie anhand von TextStorm erläutert haben, stellen stärker auf Vorschläge für die Erweiterung der Relationen einer Ontologie ab. Zudem erfordern sie eine wissensintensive Vorverarbeitung des Textkorpus. Wir ziehen daher solche Ansätze nicht weiter in Betracht, da wir eine wenig wissensintensive Verarbeitung von Textkorpora durch Kollokationen gefordert haben.
Streng betrachtet liegen nur bei den Ansätzen Hofmanns, Agirres, Yarovskys und bei Clouds Methoden zur Anreicherung bestehender Ontologien

durch Begriffsvorschläge vor. Daher müssen wir uns bei der Bewertung auch die Frage stellen, welche der im vorliegenden Kapitel dargestellten Ansätze auch leicht zu Anreicherungsansätzen modifizierbar sind.
Der Ansatz des hierarchischer Clusterbildung wird erfolgreich auf die automatische Konstruktion allgemeiner Ontologien und Thesauri angewandt. Die aktuellen experimentellen Untersuchungen verschiedener Ähnlichkeitsheuristiken wie bei Mädche et al. [68] sind ebenfalls in diesem Zusammenhang zu sehen. Obwohl diese Experimente durchaus als ähnlichkeitsbasierte Anreicherung einer bestehenden Ontologie verstanden werden können, treffen sie nicht unsere im vorigen Kapitel hergeleiteten Teilziele. Die Struktur der gegebenen Ontologie (ein englischsprachiger Thesaurus) sieht hier stets mehrere gegebene natürlichsprachliche Begriffsbezeichner pro Begriff vor. Mädche et al. [68] untersuchen unter anderem die diese Wortmengen am besten repräsentierenden Vektoren und bewegen sich damit innerhalb typischer Fragestellungen der Clusterverfahren. Die Festlegung eines Kriteriums, um Begriffe überhaupt zu Vorschlägen werden zu lassen oder die Zuordnung zu mehreren Clustern wird von den genannten Autoren nicht untersucht. Dies liegt darin begründet, dass das eigentliche Ziel der Autoren darin besteht, eine Ontologie zu konstruieren und nicht eine bestehende Ontologie anzureichern. Die vollständig automatische Konstruktion von Ontologien durch hierarchisches Clustern liefert zudem immer die offene Frage nach der Benennung einzelner Cluster innerhalb der entstandenen Hierarchie.
Die Experimente bei [68] zeigen, dass für die hierarchische Clusterbildung Attribute in Frage kommen, die die Kollokationen in einem Textkorpus auswerten. Angesichts des Zusatzwissens, das für Attribute mit syntaktischer Auswertung (beipielsweise Subjekt-Prädikat-Objekt-Strukturen), modelliert werden müsste, behalten wir die Vorgehensweise kollokationsbezogener Attribute bei.
Die Kritik an Hofmanns Ansatz lässt sich in ähnlicher Weise formulieren. Hier bestehen ebenfalls mehrere, in einigen Versuchsaufbauten sogar viele Repräsentanten der einzelnen Klassen (die Begriffen entsprechen). Mehrfachzuordnungen oder die Definition einer Rangliste bester Zuordnungen eines neuen Begriffes zu den bestehenden Begriffen würden hier eine zusätzliche Erweiterung des Ansatzes erfordern. Insgesamt bleibt festzuhalten, dass sowohl die überwachten als auch die nicht überwachten theoretischen Ansätze aus unserer bisherigen Darstellung nicht für fachgebietsspezifische Ontologien erprobt wurden.
FCA-Merge könnte zwar auf eine reine Ontologieanreicherung übertragen werden, der formal-logisch einsichtige Ansatz von Stumme liefert allerdings weder eine deutliche Unterscheidung von zusätzlichen Relations- oder Be-

griffsvorschlägen noch direkt handhabbare und integrierbare natürlichsprachliche Begriffsbezeichner. Er ist daher für unsere Zwecke nicht geeignet und wird nicht weiter betrachtet.
Rein kollokationsbasierte Ansätze operieren einerseits zwar nicht mit Ähnlichkeitsfunktionen, die auf einem Sprachmodell beruhen. Andererseits ist jedoch hervorzuheben, dass sie Mehrfachzuordnungen und auf direkte Weise Eingangskriterien für neue Begriffe (und nicht nur deren Einordnung), wenn sie (als Vektor-Repräsentation) bereits vorhanden sind, zulassen. Obwohl sie somit nicht alle unserer Anforderungen erfüllen, werden sie dennoch als naive Anreicherungsstrategie die Grundlage für qualitative Vergleiche unseres im nächsten Kapitel zu entwickelnden Verfahrens bilden.

Kapitel 5

Ontologieanreicherung

In diesem Kapitel entwickeln wir das mathematische Modell einer ähnlichkeitsbasierten Ontologieanreicherung. Wir verstehen Anreicherung hier als die Sammlung und Einordnung von Begriffsvorschlägen zu einer bestehenden fachgebietsspezifischen Ontologie.

5.0.1 Überblick über den Ansatz

Die generelle Idee des vorzustellenden Verfahrens begründet sich durch folgende Überlegungen. Da wir in einer Sammlung natürlichsprachlicher Texte für jeden Begriff aus der Ontologie seinen typischen sprachlichen Kontext, seine typische Verwendungsweise ermitteln können, ist auch ein Vergleich dieser Verwendungsarten möglich. Aus einem solchen Vergleich resultiert ein Zahlenwert, der die Ähnlichkeit oder Unähnlichkeit der typischen sprachlichen Verwendung zweier Begriffe darstellt. Gleichzeitig liefert auch die Struktur der Ontologie ein Verständnis für die Ähnlichkeit zweier Begriffe. In Analogie zu Graphen lassen sich für Paare von Begriffen aus der Ontologie Ähnlichkeiten formulieren. Für jedes Paar von Begriffen aus der Ontologie liegen somit zwei Arten des Verständnisses ihrer Ähnlichkeit vor. Wir werden ein mathematisches Verfahren entwickeln, dass diese beiden Arten der Ähnlichkeit (oder Unähnlichkeit) einander anpasst. Das Resultat ist für eine gegebene Ontologie und einen gegebenen Textkorpus eine Neudefinition von Ähnlichkeit und Unähnlichkeit, die auch auf die anderen Wörter aus dem Textkorpus angewandt werden kann. Unser Algorithmus bestimmt danach für Begriffskandidaten (Wörter oder Wortgruppen) aus dem Textkorpus die Ähnlichkeit oder Unähnlichkeit zu bereits existierenden Begriffen einer gegebenen Ontologie.
Die Ähnlichkeiten, die sich anhand der ontologischen Struktur bestimmen

lassen, liefern auch das Eingangskriterium darüber, ob ein Wort aus dem Textkorpus eine geeignete Ergänzung der vorhandenen Ontologie sein könnte. Wird für ein Wort aus dem Korpus die Ähnlichkeit zu einem gegebenen Begriff der Ontologie als hoch genug gewertet[1], so wird aus diesem Wort ein Begriffsvorschlag. Ein Begriffsvorschlag wird zu einem oder mehreren vorhandenen Begriffen als semantisch geeignete Stelle einer Erweiterung der Ontologie platziert.
Zentral für die Argumentation innerhalb des folgenden Kapitels ist die Überführung der im Sprachmodell vorliegenden Vektoren in neuartige, der jeweiligen Ontologie angepassten Ähnlichkeitsmaße oder Unähnlichkeitsmaße. Vom mathematischen Standpunkt aus betrachtet, definieren wir Gewichte für die einzelnen Attribute des Sprachmodells, indem wir ein Minimierungsproblem definieren.

5.0.2 Gliederung des Kapitels

Das Kapitel gliedert sich wie folgt. Abschnitt 5.1 liefert eine Abgrenzung gegenüber den verwandten Arbeiten, die in Kapitel 4 untersucht wurden. Abschnitt 5.2 formalisiert unser Verständnis von Ontologieanreicherungen. In Abschnitt 5.3 zeigen wir zunächst anhand einer Matrixrepräsentation die typische Art und Weise, wie die Verwendung von Wörtern aus einem Textkorpus für unser Verfahren erfasst wird (Unterabschnitt 5.3.1). Daraus ist auch ein erstes Ontologieanreicherungsverfahren (Algorithmus 1) ableitbar. Der Unterabschnitt 5.3.2 widmet sich der Anpassung von ontologischer Ähnlichkeit und der aus Korpusinformationen resultierenden Ähnlichkeit. Ein ähnlichkeitsbasiertes Verfahren (Algorithmus 2) wird daraus abgeleitet. Abschnitt 5.4 zeigt verschiedene Ausprägungen der ähnlichkeitsbasierten Ontologieanreicherung, unter anderem auch als unähnlichkeitsbasierte Ontologieanreicherung (Algorithmus 3 als Variante von Algorithmus 2). Die Ausprägungen hängen mit verschiedenen Formulierungen ontologischer und vektorwertiger Vergleichsmaße zusammen. Der Abschnitt 5.5 zeigt automatische Evaluationsverfahren für Ontologieanreicherungsverfahren. Wir schließen das Kapitel mit einer Zusammenfassung der eigenen Beiträge des vorliegenden Kapitels (Abschnitt 5.6).

[1] Oder in Analogie: wenn die Unähnlichkeit als niedrig genug befunden wird.

5.1 Abgrenzung

Die im letzten Kapitel erläuterten Ansätze weisen einige spezifische Vorteile auf, die wir übernehmen werden. Der im Folgenden zu entwickelnde Ansatz benutzt ein Sprachmodell mit einer Attributmenge $\mathcal{A}$. Dies liegt darin begründet, dass wir damit Ähnlichkeitsfunktionen definieren können. Wir legen ähnlich wie bei Clusterverfahren für jedes Paar von Vektoren, die sich zu einem Paar von Wörtern aus einem gegebenen Textkorpus durch Abgleich mit $\mathcal{A}$ ergeben, Ähnlichkeitsfunktionen fest. Die Ähnlichkeitsfunktionen benötigen wir, um die im Kapitel zu technischen Ontologieerstellungsprozessen gefundenen Forderungen an einen automatisch unterstützten Ablauf der Ontologiekonstruktion erfüllen zu können. Das betrifft insbesondere die Varianten der Zuordnung; sowohl die beste Zuordnung eines neuen Begriffes als auch Mehrfachzuordnungen oder Verbesserungsvorschläge von Zuordnungen können durch die numerische Definition einer Ähnlichkeitsfunktion gesteuert werden, indem man die Wörterpaaren zugehörigen Ähnlichkeitswerte ordnet.
Für unsere Definition der ähnlichkeitsbasierten Ontologieanreicherung sind im Vergleich zu den verwandten Arbeiten folgende Anpassungen nötig. Im Gegensatz zu den nicht überwachten Methoden aus dem vorigen Kapitel werden wir gegebene Ontologien mit mindestens einem natürlichsprachlichen Bezeichner pro Begriff als zusätzliche Informationsquelle für neue Ähnlichkeitsmaße nutzen. Dies ist insbesondere für den Kern eines Ontologieanreicherungsverfahrens, das Ontologieautoren Begriffsvorschläge präsentieren soll, notwendig. Die Ähnlichkeitswerte der gegebenen Ontologie liefern auch ein Eingangskriterium darüber, ob ein Begriff überhaupt vorgeschlagen werden sollte. Dieser Zusammenhang wird sich in unserer Definition von Schranken in diesem Kapitel niederschlagen. Überschreitet ein Ähnlichkeitswert zwischen einem gegebenen und einem potentiellen Begriff eine Schranke, so wird der potentielle Begriff zum Vorschlag für den gegebenen Begriff. Schranken wirken somit prinzipiell anders als Cluster- oder Kategorisierungsverfahren, bei denen jeweils auch mit geringen Ähnlichkeitswerten eine Zuordnung erfolgen kann.

5.2 Formalisierung der Ontologieanreicherung

Der vorliegende Abschnitt liefert eine allgemeine Formalisierung der Ontologieanreicherung durch neue Begriffe.
Im Gegensatz zu den hierarchischen Clusterverfahren Abschnitt 4.1.1 basiert

die Ontologieanreicherung, wie wir sie im vorliegenden Abschnitt definieren, stets auf gegebenen Ontologien. Zudem sollen die Ontologieanreicherungsverfahren die kooperative Ontologieanreicherung unterstützen. Daher grenzen wir Ontologienareicherungsverfahren als diejenigen Verfahren ein, die

- für eine gegebene Ontologie Ω Begriffsvorschläge erzeugen
- jedem Begriffsvorschlag einen Platz in Ω zuweisen
- die aus einem Textkorpus mittels eines Sprachmodells erhobenen statistischen Daten zur Verwendung des Begriffsbezeichners als Eingangsgrößen verwenden

Diese Charakterisierung weist gegenüber hierarchischen Clusterverfahren dann einen höheren Realitätsbezug auf, wenn wir davon ausgehen, dass die Autoren fachgebietsspezifischer Ontologien einen Begriff als hinreichend durch seinen natürlichsprachlichen Bezeichner beschrieben ansehen. Das Sammeln mehrerer Bezeichner in Form von Synonymlisten, welche als Augangspunkt eines hierarchischen Clusterverfahrens (und darauf basierender Anreicherungsverfahren) verstanden werden könnten, stellt in unserem Anwendungsbezug eher einen Ausnahmefall dar. Wir benutzen daher die Zusatzforderung zur Definition 4 als Minimalanforderung für ein Ontologieanreicherungsverfahren: pro Begriff existiert zunächst genau ein natürlichsprachlicher Bezeichner. Ein Ontologieanreicherungsverfahren muss dementsprechend gestaltet werden.
Die Formalisierung unseres Verständnisses von Ontologieanreicherungen findet sich in folgender Definition.

Definition 8 (Ontologieanreicherung) *Sei ξ ein Textkorpus, sei C eine Menge von Wörtern oder Wortgruppen aus ξ. Ein Ontologieanreicherungsalgorithmus ist ein Algorithmus, der ξ und eine gegebene Ontologie $\Omega := \{B, \leq, R, \sigma\}$ als Eingangsdaten benutzt und für alle $b \in B$ eine Menge $P(b) \subseteq C$ als Ergebnis liefert. Wir nennen $P(b)$ die Menge der Begriffsvorschläge für b. Die Menge C mit $C \supseteq \bigcup_{b \in B} P(b)$ nennen wir Kandidaten.*

Die Menge der Kandidaten kann vordefiniert sein, beispielsweise als Wörter, die sich im Index eines Fachbuches befinden oder als Sammlung von Wörtern, die noch nicht als Begriffe in die Ontologie aufgenommen wurden, für eine Anwendung der Ontologie zur Verschlagwortung aber notwendig sind. Die Kandidaten können aber auch freier vorgegeben sein, wie etwa als Menge aller Wörter und Wörtgruppen, die sich im Textkorpus ξ befinden. Anhand

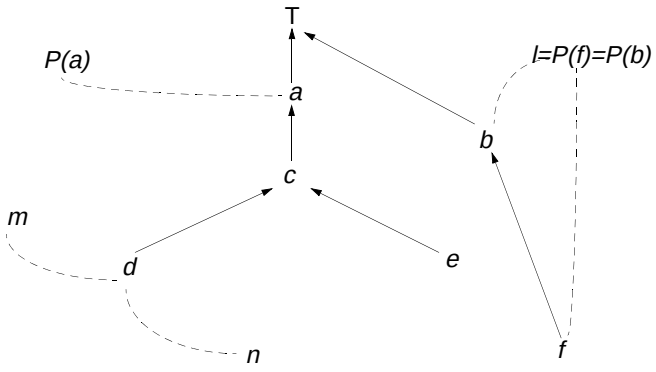

Abbildung 5.1: Schematisches Beispiel für Ontologieanreicherungen

der Abbildung 5.1 zeigen wir ein Beispiel der Ontologieanreicherung. Wiederum sind hier T, a, b, c, d, e, f Begriffe aus einer gegebenen Ontologie. Die gerichteten Kanten entsprechen der Unterbegriffsrelation, die gestrichelten Linien verbinden Begriffe und ihre Begriffsvorschläge. Die Grafik zeigt verschiedene Situationen, die bei Begriffsvorschlägen entstehen können. $P(a)$ sind die Begriffsvorschläge für a. Über die Mächtigkeit $|P(a)|$ findet sich zunächst keinerlei Aussage in der Grafik. Für die Begriffsvorschläge zu den Begriffen b, f und d wird dies expliziter. Es gilt $\{m, n\} = P(d)$, was zeigt, dass zwei Begriffsvorschläge m und n zu einem Begriff d erzeugt werden können. Außerdem sehen wir in der Grafik die Identität $\{l\} = P(b) = P(f)$. Dies stellt ein Beispiel für die Zuordnung eines Begriffsvorschlages l zu mehreren Begriffen der Ontologie dar.

5.3 Ontologieanreicherungsverfahren

Als nächsten Grundbestandteil des Ansatzes werden wir in 5.3.1 eine formale Repräsentation der Begriffe aus einer gegebenen Ontologie Ω definieren. In 5.3.2 führen wir Gewichtungen ein, die durch die Verstärkung oder Minde-

rung des Einflusses einzelner Attribute aus einem Sprachmodell Vergleiche zwischen den Kollokationsinformationen beeinflussen können. Eine optimale Gewichtung passt ontologische und vektorwertige Vergleichsmaße einander an (5.3.2.2). Die optimale Gewichtung ist insofern auch für jede Ontologie und mit einem dazugehörigen Textkorpus eine Neudefinition von Ähnlichkeit. Diese Neudefinition bildet die Grundlage der allgemeinen Formulierung ähnlichkeitsbasierter Ontologieanreicherung (5.3.2.3).

5.3.1 Matrixrepräsentation

Die Matrixschreibweise bringt den Vorteil mit sich, dass wir im Gange unserer Arbeit verschiedenartige Formulierungen des Algorithmus mit Hilfe einer Referenznotation, der Matrix, darstellen können.
Wir definieren $M(\xi, \Omega, \mathcal{A})$ als Matrix für eine endliche Menge aus Begriffsbezeichnern, eine endliche Menge von Attributen $\mathcal{A}$ eines Sprachmodells und einem gegebenen Korpus ξ. Wir widmen uns nun den Matrixeinträgen.
Ähnlich zu den Arbeiten in der hierarchischen und allgemeinen Clusterbildung aus Kapitel 4 definieren wir hier Matrixeinträge über Vektoren. Jedes Wort und jede Wortgruppe aus ξ kann als Vektor dargestellt werden. Der Vektoreintrag zeigt an, wie oft das i-te Attribut aus $\mathcal{A}$ erfüllt wurde. Jede Zeile der Repräsentationsmatrix entspricht dem Vektor für den Begriffsbezeichner und wird auf Grundlage der Auswertung von ξ erstellt. Wir nennen die Zeilen der Repräsentationsmatrix auch *Begriffsvektoren*. Um über das generelle Phänomen dünn besetzter Vektoren bei der Erfassung durch Sprachmodelle [59] hinaus vollständig leere (nullwertige) Einträge in $M(\xi, \Omega, \mathcal{A})$ zu vermeiden, sollte ξ die Begriffsbezeichner $D(\Omega)$ enthalten.
Unsere Grundannahme bleibt, dass wir Begriffsvorschläge als natürlichsprachliche Bezeichner anbringen möchten. Die weiteren Nomen, Adjektive, Verben und Wortgruppen aus ξ identifizieren wir in der Folge mit $B(\xi)$. $D(\Omega)$ stehen somit für die bereits bekannten Begriffe aus der Ontologie, wohingegen $B(\xi)$ als bislang unbekannte Begriffe aufgefasst werden können, und es gilt $D(\Omega) \cap B(\xi) = \emptyset$.
Wie wir in unserer Kritik der verwandten Ansätze (Abschnitt 4.2) bereits erklärten, operieren wir mit kollokationsbezogenen Attributen. Diese erfordern einen geringen Zusatzaufwand bei der Sprachverarbeitung und wurden in den empirischen Arbeiten von [37] und [68] als prinzipielle Möglichkeit für hierarchische Clusterverfahren identifiziert.
Wir konstruieren nun ein Beispiel für eine Repräsentationsmatrix zu einem Korpus ξ_b, der aus zwei natürlichsprachlichen Beispielsätzen bestehen wird. Dies dient lediglich der Veranschaulichung und stellt einen unrealistisch klei-

nen Korpus dar. Die detaillierte Darstellung dieses Unterabschnittes wurde deshalb gewählt, weil wir bei der Implementierung einer Testumgebung auf die Struktur der Repräsentationsmatrix zurückgreifen werden.
Gegeben sei zunächst ein Satz, der den Textkorpus bildet:

'Der Bäcker backt frisches Brot'

Wenn wir mit δ_W den in Wörtern gezählten Abstand zweier Wörter eines Wortpaares bezeichnen, so ergeben sich für den Beispielsatz die Werte der folgenden Tabelle 5.1. Um für das Beispiel weiterhin die Attribute in $\mathcal{A}_b$ zu

Tabelle 5.1: Wortabstände δ_W im Satz 'Der Bäcker backt frisches Brot'

$Wort$	$Wort$	δ_W
$Bäcker$	$Brot$	3
$Bäcker$	$frisches$	2
$frisches$	$Brot$	1
$Brot$	$Brot$	0

konstruieren, wählen wir nun ein festes δ_W. Wenn b_j der j-te Bezeichner aus $D(\Omega)$ ist, dann sei das j-te Attribut aus $\mathcal{A}$ dadurch erklärt, dass in ξ_b mit b_j in einem Maximalabstand δ_W eine Kollokation vorliegt. Die Einträge m_{ij} in der Repräsentationsmatrix $M(\xi, \Omega, \mathcal{A})$ ergeben sich folgendermaßen:

$$m_{ij} = 0, \tag{5.1}$$

wenn $b_i \in D(\Omega)$ das Attribut $\mathcal{A}_j \in \mathcal{A}$ nicht erfüllt und

$$m_{ij} = M \tag{5.2}$$

wenn $b_i \in D(\Omega)$ das Attribut $\mathcal{A}_j \in \mathcal{A}$ genau M mal erfüllt.
Wir kommen nun zur eigentlichen Konstruktion der Repräsentationsmatrix. Es sei $D(\Omega) = \{Butter,\ Brot,\ Frühstück\} = B$. Wir nehmen also wiederum an, dass sich Begriffe und Bezeichner entsprechen. ξ_b bestehe aus zwei Sätzen:

'Der Bäcker backt frisches Brot. Brot und Butter gehören zu jedem anständigen Frühstück'.

Mit einer aus dem Wortabstand $\delta_W = 3$ konstruierten Menge von Kollokationsattributen $\mathcal{A}_b := \{\mathcal{A}_1, \mathcal{A}_2\, \mathcal{A}_3\}$ - wobei $\mathcal{A}_1$ die Kollokation mit 'Bäcker',

A_2 die Kollokation mit 'Brot' und $\mathcal{A}_3$ die Kollokation mit 'Frühstück' kennzeichnet - erhalten wir folgende Matrix:

	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$
Butter	0	2	0
Brot	1	2	0
Frühstück	0	0	1

Die Kollokation kann hier auch über die Satzgrenze hinausgehen. Die Schreibweise δ_W für Wortabstände im Textkorpus können wir verwenden, um das bereits im letzten Kapitel aufgeführte kollokationsbasierte Anreicherungsverfahren im Pseudokode (Algorithmus 1) darzustellen.

Data : eine Ontologie $\Omega := (B, \leq, R, \sigma)$, ein Textkorpus ξ, eine natürliche Zahl δ_W als Wortabstand, Kandidaten C
Result : Begriffsvorschläge $P(b)$ für alle $b \in B \backslash \{\top\}$
for *alle* $b \in B \backslash \{\top\}$ **do**
  $P(b) \leftarrow \varnothing$
  for *alle* $c \in C$ **do**
    if *c Kollokator im Abstand von maximal* δ_W *zum natürlichsprachlichen Bezeichner von b in* ξ **then**
      $P(b) \leftarrow P(b) \cup c$
    end
  end
end

Algorithmus 1 : Naive Ontologieanreicherung

Wir nennen die in Algorithmus 1 dargestellte Art der Anreicherung naiv, da sie weder eine sprachliche Vorverarbeitung noch Ähnlichkeitsfunktionen beinhaltet und lediglich auf das gemeinsame Vorkommnis von Wörtern abstellt. Beispielsweise müssen Synonyme eines Begriffsbezeichners explizit im Textkorpus in der Nähe eines Begriffsbezeichners auftreten, um zum Begriffsvorschlag zu werden. Wenn die Wortverwendung betrachtet wird (wie beispielsweise beim Clustern), so lassen sich Synonyme auch implizit über einen ähnlichen sprachlichen Kontext finden.
Im Gegensatz zum rein kollokationsbasierten naiven Anreicherungsverfahren steht die im Folgenden zu entwickelnde ähnlichkeitsbasierte Anreicherung.

Die Form der obigen Repräsentationsmatrix wird in den nun folgenden Unterabschnitten die Grundlage einer Transformation, die den paarweisen Vergleich der Zeilen (den kollokationsbasierten Vergleich der Wortverwendungen in einem Textkorpus) an ontologische Ähnlichkeits- und Abstandsmaße anpasst. Die Anpassung wiederum erfolgt durch Skalarmultiplikationen der Spalten. Die eigentlichen Zielgrößen der Transformation stellen wir im folgenden Abschnitt ausführlich dar. Die prinzipielle Umsetzung des Textkorpus zu einem auf Kollokationen und einem Wortabstand δ_W beruhenden Sprachmodell und einer wiederum darauf basierenden Repräsentationsmatrix behalten wir bei, da diese Art der Attribute auch ohne grammatikalische und syntaktische Zusatzbetrachtungen eine Erfassung charakteristischer Eigenschaften der gegebenen Begriffsbezeichner im Textkorpus möglich werden lässt.

5.3.2 Gewichtungen

Der folgende Abschnitt erklärt, wie die Gewichtung einzelner Kollokationsattribute die Ähnlichkeitsbestimmung zweier Vektoren als Repräsentanten von Wörtern und Begriffen beeinflusst. Gewichtungen sind somit für die Definition von Ähnlichkeitsmaßen bei der Ontologieanreicherung von entscheidender Bedeutung.
Ähnlich wie im hierarchischen Clusterverfahren nach Bisson, Nedellec und Canamero [10] besteht unser Ansatz darin, jedes Attribut aus $\mathcal{A}$ mit einer Gewichtung zu versehen. [10] wählte einen Ansatz mit Gewichtungen, um über eine Variation dieser Gewichtungen diejenigen zu finden, die eine übermäßige Verallgemeinerung der Cluster vermied. Unsere Zielrichtung unterscheidet sich hiervon; wir arbeiten mit der Grundannahme, dass sich die Struktur der gegebenen Ontologie eines Ontologieanreicherungproblems in Ähnlichkeiten oder Unähnlichkeiten übersetzen lässt. Wir versuchen, dies mit Hilfe der Gewichtung, technisch gesprochen mittels der Skalarmultiplikation einzelner Spalten aus einer Repräsentationsmatrix $M(\xi, \Omega, \mathcal{A})$, nachzubilden.

In der Abbildung 5.2 wird die Motivation der Gewichtungen vermittelt. Zu den einzelnen Begriffen der gegebenen Ontologie wurden auf Grundlage einer gegebenen Attributmenge $\mathcal{A}$ mit $|\mathcal{A}| = n$ die zum Begriff gehörigen Zeilen der Repräsentationsmatrix erzeugt. Diese wiederum sind komponentenweise mit Gewichten $k_1, ..., k_n$ multipliziert, so dass jeder Abstand der Ontologie durch eine Wahl von $k_1, ..., k_n$ angepasst werden kann. Dies ist durch die gestrichelte dünne Linie zwischen den zu *Bakterium* und *Salmonella* versinnbildlicht: wenn wir den zu den dicker dargestellten Kanten gehörigen

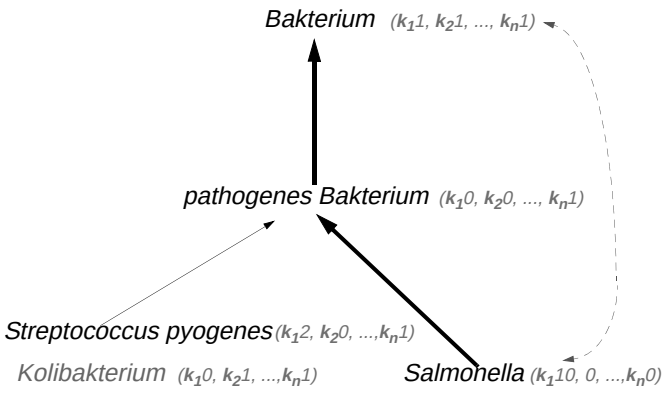

Abbildung 5.2: Motivation der Gewichtungen

ontologischen Zusammenhang '*Salmonella* ist ein Unterbegriff von *Bakterium*, allerdings kein direkter Unterbegriff' auch beim Vergleich der Vektoren (Zeilen der Repräsentationsmatrix) $(k_1 \cdot 10, ..., k_n \cdot 0)$ und $(k_1 \cdot 1, k_2 \cdot 1, ..., k_n \cdot 1)$ durch einen erhöhten oder niedrigen Wert des Vergleiches berücksichtigen wollen, so geschieht das über konkrete Werte $k_1, ..., k_n$.

5.3.2.1 Gewichtungen und Vergleichsmaße

Wir definieren Gewichtungen als zentrale Einflussgrößen für das Anreicherungsproblem. Diese Gewichtungen werden sodann in ein Minimierungsproblem eingehen, über das eine Anpassung der Korpusinformationen an die Informationen aus der Ontologiestruktur selbst erreicht wird.

Definition 9 (Gewichtung) *Sei $\mathcal{A}$ eine Menge von Kollokationsattributen mit $|\mathcal{A}| = n$. Ein Vektor $\vec{k} = (k_1, ..., k_n)$ reeller Zahlen mit $\vec{k} \in \mathbb{R}^n$ heißt Gewichtung von $\mathcal{A}$.*

Insbesondere wird die Gewichtung $\vec{k}$ in die Bewertung neuer Begriffe aus der Menge der Kandidaten C (wie in der Abbildung 5.2 *Kolibakterium*)

eingehen. Die Skalarmultiplikation zur Anwendung der Gewichtung, wie sie in der Abbildung 5.2 angedeutet wurde, kann auch auf das (vektorwertige) Resultat eines Vergleiches zweier Zeilen der Repräsentationsmatrix angewandt werden. Wir werden in den Folgeabschnitten sehen, dass diese leicht von der Darstellung im Beispiel in Abbildung 5.2 abweichende Methode mit verwandten Arbeiten und Maßen der Computerlinguistik vereinbar ist. Insgesamt jedoch halten wir fest, dass in unserem Ansatz nach den Definitionen 8 und 9 für die Anreicherung einer gegebenen Ontologie $\Omega := (B, \leq, R, \sigma)$ festgelegt wird:

$$\exists \vec{k} \in \mathbb{R}^n \forall b \in B : P(b) := P(b, \vec{k}) \tag{5.3}$$

Die Grundlage hierfür schaffen wir durch ein vektorwertiges Vergleichsmaß, das für Paare von Begriffsvektoren existieren muss.

Definition 10 (vektorwertiges Vergleichsmaß) *Eine aus* $\Omega := (B, \leq, R, \sigma)$, ξ *und* $\mathcal{A}$ *erklärte Abbildung* $S_v : B \times B \rightarrow (\mathbb{R}^+)^{|\mathcal{A}|}$ *heißt vektorwertiges Vergleichsmaß. Ein vektorwertiges Vergleichsmaß bildet Paare von Zeilen der Repräsentationsmatrix* $M(\Omega, \xi, \mathcal{A})$ *auf Vektoren der Länge* $|\mathcal{A}|$ *ab.*

Wir definieren nun die Zielgröße, die durch die Gewichtung erreicht werden soll.

Definition 11 (ontologisches Vergleichsmaß) *Eine aus* $\Omega := (B, \leq, R, \sigma)$ *allein erklärte Abbildung* $S_\Omega : B \times B \rightarrow \mathbb{R}_0^+$ *heißt ontologisches Vergleichsmaß.*

Spezielle Fälle dieser ontologischen Vergleichsmaße sind Ähnlichkeitsmaße und Unähnlichkeitsmaße. Ein Ähnlichkeitsmaß wird in der Regel für 'nahe beieinanderliegende', das heißt nur über einen kurzen Pfad weniger Relationen aus $R \cup \{\leq\}$ wechselseitig erreichbare Begriffe hohe Werte liefern, ein Unähnlichkeitsmaß wirkt umgekehrt[1]. Eine sehr einfache Heuristik würde in der Situation der Abbildung 5.2 die Länge (2 Schritte) des dick visualisierten Pfades verwenden. Wir werden in den Folgeabschnitten auch hierzu mehrere Heuristiken vorstellen.

5.3.2.2 Minimierungsproblem

Wir sind nun dazu in der Lage, unser zentrales Anpassungsproblem, nämlich die erwünschte Übereinstimmung des vektorwertigen und des ontologischen

[1] Wir sprechen von Unähnlichkeit, nicht von Distanz, da wir die metrischen Axiome von Distanzmaßen nicht verifizieren. Unähnlichkeitsmaße dienen als Gegenstück der Ähnlichkeitsmaße.

Vergleichsmaßes für einen gegebenen Korpus und eine gegebene Ontologie, als Minimierungsproblem zu formulieren.
Seien dazu ein vektorwertiges Vergleichsmaß S_v und ein ontologisches Vergleichsmaß S_Ω gegeben. Über eine Kleinste-Quadrate-Anpassung formulieren wir das Ziel einer möglichst großen Übereinstimmung der beiden Vergleichsmaße. Wir suchen das Minimum der Summe:

$$\min_{\vec{k}\in\mathbb{R}^{|\mathcal{A}|}} \Big(\sum_{(b_1,b_2)\in(B\setminus\{\top\})\times(B\setminus\{\top\})} (S_\Omega(b_1,b_2) - \vec{k}S_v(b_1,b_2))^2 \Big) \tag{5.4}$$

wobei mit B wiederum die Begriffe einer gegebenen Ontologie Ω gemeint sind und mit $\vec{k}S_v(b_1,b_2)$ das Skalarprodukt aus einer Gewichtung und einem vektorwertigen Vergleichsmaß. Sobald wir das Minimierungsproblem 5.4 durch ein $\vec{k}_{opt}$ lösen, erhalten wir eine an die gegebene Ontologie angepasste Abbildung von $B \times B$ in die reellen Zahlen und somit ein neues ontologisches Vergleichsmaß. Die Lösung $\vec{k}_{opt}$ des Minimierungsproblems nennen wir auch **optimale Gewichtung**. Das resultierende ontologische Vergleichsmaß hat die Eigenschaft, dass es durch eine Skalarmultiplikation der optimalen Gewichtung mit den vektorwertigen Resultaten auf beliebige Paare von Wörtern aus ξ ausdehnbar ist. Als Abbildung s formuliert erhalten wir:

$$s : B(\xi) \times B(\xi) \to \mathbb{R} \tag{5.5}$$

mit

$$s(x,y) := \vec{k}_{opt}S_v(x,y) \tag{5.6}$$

Der folgende Unterabschnitt bringt dies in Zusammenhang mit der Ontologieanreicherung. Zuvor möchten wir noch ein kleines Beispiel einführen, das zeigt, wie die Repräsentationsmatrix aus 5.3.1 mittels der Gewichtung beim Minimierungsproblem und bei der Ontologieanreicherung eingesetzt wird. Es seien *Butter* und *Brot* Begriffe einer Ontologie mit *Butter* $\leq \top$, *Brot* $\leq \top$ und *Frühstück* $\leq \top$. Es werden über ein vektorwertiges Vergleichsmaß die Vektorpaare *Butter* und *Brot*, *Butter* und *Frühstück* und *Frühstück* und *Brot* verglichen. Der Begriffsvektor zu *Butter* ist $(0,2,0)'$, der Begriffsvektor zu *Brot* $(1,2,0)'$. Anhand der in den folgenden Abschnitten noch herzuleitenden Rechenregeln dafür könnte daraus beispielsweise der Vektor $(0,2,0)'$ resultieren (komponentenweise Minimabildung). Eine optimale Gewichtung $(k_{opt,1}, k_{opt,2}, k_{opt,3})$ ist dafür verantwortlich, über die Skalarmultiplikation $(k_{opt,1}, k_{opt,2}, k_{opt,3})(0,2,0)' = 2k_{opt,2}$, die Ähnlichkeit

(zum Beispiel die inverse Pfadlänge $2^{-1} = 0.5$, die sich entlang $\leq$ und $\geq$ ergäbe) von *Butter* und *Brot* in der gegebenen Ontologie herzustellen. Das Minimierungsproblem besteht in diesem Fall aus neun Summanden, einer davon wird $(2k_{opt,2} - 0.5)^2$ sein. Ist dann $\vec{k}_{opt}$ gefunden, so kann für jeden Vektor zu einem Wort aus dem Textkorpus ξ_b eine Ähnlichkeit zu jedem gegebenen Begriff berechnet werden. Wenn der Kandidat *Margarine* beispielsweise durch den Vektor $(0, 1, 0)'$ repräsentiert wäre, so ergäbe sich aus komponentenweiser Minimabildung und Skalarmultiplikation $k_{opt,2}$ als Ähnlichkeitswert mit *Butter*. Ist dieser Wert $k_{opt,2}$ höher als die Schranke, die man für Begriffsvorschläge zu *Butter* festlegt (beispielsweise 1, weil dies die Länge einer Kante entlang $\leq$ ist), so wird aus *Margarine* ein Begriffsvorschlag, der zu *Butter* platziert wird.

5.3.2.3 Anreicherung mit optimalen Gewichtungen und Schranken

Wir können nun die ähnlichkeitsbasierte Ontologieanreicherung unter Zuhilfenahme von Gewichtungen sowie ontologischen und vektorwertigen Vergleichsmaßen definieren.
In Anlehnung an Definition 8 spezifizieren wir für einen gegebenen Begriff b, eine Kandidatenmenge C und ein $T_b \in \mathbb{R}$ die Menge der Begriffsvorschläge $P(b)$ als

$$P(b) := \{x \in C | \vec{k}_{opt} S_v(x, b) \geq T_b\}. \tag{5.7}$$

Dies bedeutet, dass für den Kandidaten x die durch $\vec{k}_{opt} S_v(x, y)$ gegebenen neu definierten Ähnlichkeiten eine bestimmte *Schranke* T_b überschreiten muss, um zum Begriffsvorschlag zu werden. Für T_b kommen per Konstruktion unseres ähnlichkeitsbasierten Ansatzes Werte in Frage, die im Zusammenhang mit dem ontologischen Vergleichsmaß stehen, für das k_{opt} gefunden wurde. Eine mögliche Heuristik, von der wir Gebrauch machen werden, ist die Berechnung von T_b aus dem ontologischen Vergleichsmaß durch eine Durchschnittsbetrachtung. Dazu definieren wir eine Menge X. Die Elemente aus wären in unserer üblichen Visualisierung einer Ontologie (beispielsweise in 2.4) alle genau eine zur Relation $\leq$ gehörige Kante von b entfernt. Somit erhalten wir $X :=$
$\{x \in B | b \leq x, \{y | b \leq y \leq x\} = \emptyset\} \cup \{x \in B | b \geq x, \{y | b \geq y \geq x\} = \emptyset\}$. Für T_b definieren wir:

$$T_b := \frac{\sum_{x \in X} S_\Omega(b, x)}{|X|} \tag{5.8}$$

Prinzipiell wird eine Variation von T_b die Anzahl der Vorschläge und somit tendenziell auch auf die Genauigkeit des Ergebnisses beeinflussen.
Im Quelltext Algorithmus 2 folgt eine Darstellung des Anreicherungsverfahrens als Pseudokode.

Data : eine Ontologie $\Omega := (B, \leq, R, \sigma)$, eine optimale Gewichtung $\vec{k}_{opt}$, ein Textkorpus ξ, ein vektorwertiges Ähnlichkeitsmaß S_v, eine reellwertige Schranke T_b pro Begriff, Kandidaten C
Result : Begriffsvorschläge $P(b)$ für alle $b \in B \backslash \{\top\}$
for *alle* $b \in B \backslash \{\top\}$ **do**
  $P(b) \leftarrow \varnothing$
  for *alle* $c \in C$ **do**
    $S_c \leftarrow \vec{k}_{opt} S_v(b, c)$
    if $S_c > T_b$ **then**
      $P(b) \leftarrow P(b) \cup \{c\}$
    end
  end
end

Algorithmus 2 : Ähnlichkeitsbasierte Ontologieanreicherung

Der folgende Abschnitt widmet sich den konkreten Alternativen, die bei der Berechnung der ontologischen und der vektorbasierten Vergleichsmaße in Frage kommen. Die Ergebnisse werden stets mit Hilfe der Repräsentationsmatrix ausformuliert.

5.4 Ausprägungen der Anreicherung

Der folgende Abschnitt zeigt auf, welche Varianten und Definitionen sich für die ähnlichkeitsbasierte Ontologieanreicherung im Detail ergeben. Die Vielzahl der Varianten stammt dabei aus Kombinationen unterschiedlicher in der Literatur untersuchter Ansatzpunkte zur exakten Definition ontologischer und vektorwertiger Ähnlichkeitsmaße.
Gemäß des in 5.4 aufgestellten Minimierungsproblems ergeben sich zwei Ansatzpunkte zur weiteren Ausgestaltung der Anreicherung, nämlich die vektorwertigen Vergleichsmaße und die ontologischen Vergleichsmaße aus Definition 10 und Definition 11. In beiden Fällen kann sich die Berechnung der einzelnen Vergleichsmaße auf die Stärke der Ähnlichkeit (zweier

Vektoren oder Begriffe) oder auf die Unähnlichkeit (zweier Vektoren oder Begriffe) beziehen. Kombinationen eines ähnlichkeitsbasierten vektorwertigen Vergleichsmaßes mit einem ebenfalls ähnlichkeitsbasierten ontologischen Vergleichsmaß lassen eine exakte Anwendung des Ansatzes aus dem vorigen Abschnitt zu. Alle Kombinationen, bei denen ein unähnlichkeitsbasiertes Vergleichsmaß zum Tragen kommt, lassen zwar das zentrale Minimierungs- und Anpassungsproblem 5.4 in seiner Formulierung unberührt, erfordern aber eine Modifikation des Anreicherungsschrittes, wie er in Algorithmus 2 gezeigt wurde. Zudem wird bei Kombinationen, die ein ursprünglich ähnlichkeitsbasiertes ontologisches Vergleichsmaß und ein ursprünglich unähnlichkeitsbasiertes vektorwertiges Vergleichsmaß verwenden, eine einheitliche Erfassung ihres Aussagegehaltes nötig.
Der nun folgende Abschnitt widmet sich verschiedenen ontologischen Vergleichsmaßen, von denen zwei aus verwandten Arbeiten übernommen werden konnten und eines aufgrund zusätzlicher Heuristiken definiert wurde:

- das Maß nach Resnik und das Maß nach Li wurden in den jeweiligen Arbeiten empirisch überprüft [63]. [84]. Sie unterscheiden sich durch die Tatsache, dass Li rein graphentheoretisch begründete Maße entwickelt, während Resnik auch Korpusinformationen mit einbezieht.

- sowohl Li als auch Resnik vernachlässigen bestimmte Strukturmerkmale, die bei Ontologien auftreten und die Ähnlichkeit beeinflussen können. Daher werden wir diesen im asymmetrischen Vergleichsmaß Rechnung tragen.

- Wir halten die Untersuchung eines asymmetrischen Vergleichsmaßes auch deshalb für sinnvoll, weil einige vektorwertige Vergleichsmaße asymmetrisch sind und somit Grund zur Annahme besteht, dass Asymmetrie die Berechnungen beeinflusst.

Für jedes der ähnlichkeitsbasierten ontologischen Vergleichsmaße muss nun auch eine Formulierung der Unähnlichkeit nach den im ontologischen Vergleichsmaß zum Tragen kommenden Prinzipien gefunden werden.

5.4.1 Ontologische Vergleichsmaße

Sämtliche der im Folgenden darzustellenden ontologischen Vergleichsmaße werden lediglich von den Begriffen B und der Unterbegriffsrelation $\leq$ einer gegebenen Ontologie Gebrauch machen. Die Auswertung anderer in R gegebener Relationen ist zwar denkbar, erschwert aber im Einzelfall die zügige

Berechnung der durch das ontologische Vergleichsmaß festzustellenden Ähnlichkeiten. Der Grund dafür liegt darin, dass eine Unterscheidung zwischen Relationen, die die Ähnlichkeit erhöhen und zwischen Relationen, die eine niedrigere Ähnlichkeit nach sich ziehen, getroffen werden muss. Beispielsweise kann die Ursache-Wirkungsbeziehung in der Medizin für die Ziele eines Ontologieanreicherungsalgorithmus Ähnlichkeiten erhöhen, während eine Relation wie 'ist das Gegenteil von' nicht mit einer Erhöhung der Ähnlichkeit zweier durch sie verbundener Begriffe einhergehen kann. Wir halten fest, dass wir zur Berechnung von Ähnlichkeiten und Unähnlichkeiten mittels ontologischer Vergleichsmaße lediglich die Begriffe und die Relationen $\leq$ ('ist Unterbegriff von') beziehungsweise $\geq$ ('ist Oberbegriff von') heranziehen.
Wir bewegen uns damit im Rahmen der psycholinguistischen Arbeiten von Tversky [100], der Ähnlichkeitsdefinitionen für Begriffe betrachtet. Auch Arbeiten aus dem Bereich der formalen Begriffsanalyse definieren begriffliche Ähnlichkeitsmaße mit Hilfe der (formalen) Ober- und Unterbegriffsstruktur [62].
Resnik [84] definiert in seinen Arbeiten zur Ähnlichkeit von Wörtern in der linguistischen Ontologie WordNet[28] ein ontologisches Vergleichsmaß, dass ohne Weiteres auf Ontologien mit nur einem natürlichsprachlichen Bezeichner pro Begriff übertragbar ist.

5.4.1.1 Resniks ontologisches Vergleichsmaß

Resnik zieht bei seinem Maß, das wir im Folgenden mit S^0_Ω bezeichnen wollen, den so genannten Informationsgehaltes eines Begriffes (beziehungsweise seines natürlichsprachlichen Bezeichners) heran. Resnik bezieht sich bei seinem Ansatz nicht nur auf die Struktur der Ontologie, sondern auch auf die Auftrittshäufigkeiten der Begriffsbezeichner in einem Textkorpus. Aus Arbeiten zur allgemeinen Informationstheorie (vergleiche etwa [66]) folgert Resnik, dass ein seltenes Auftreten eines Wortes w im Textkorpus den Informationsgehalt dieses Wortes erhöht. Sei dazu $count(w)$ die Gesamtanzahl des Auftretens eines Wortes oder einer Zeichenkette w in einem gegebenen Textkorpus, der insgesamt N Wörter enthält. Hierbei soll N als Summe aller Satzlängen, nicht als Umfang des Vokabulars im Korpus gelten. Der Informationsgehalt für *Wörter* lässt sich demnach als

$$-\ln[count(w)N^{-1}] \tag{5.9}$$

angeben. Um diesen Informationsgehalt auf Ontologien anwenden zu können, trifft Resnik die Modellannahme, dass sich der Informationsgehalt für einen

gegebenen Begriff aus einer Ontologie aus dem jeweiligen Informationsgehalt der Begriffsbezeichner seiner Unterbegriffe und aus seinem eigenen Informationsgehalt zusammensetzt. Resnik erklärt, dass sich durch die Verrechnung der Auftrittswahrscheinlichkeiten für einzelne Wörter eine vom speziellen zum allgemeinen Begriff monoton wachsende Wahrscheinlichkeitsfunktion ergibt. Wir wählen daher statt N ein N', das sich aus den Häufigkeiten der Begriffsbezeichner zusammensetzt:

$$N' := |B|^{-1} \sum_{b_i \in B \setminus \top} count(D(b)) + \sum_{b_i \in B \setminus \top} count(D(b)) \tag{5.10}$$

Der erste Summand liefert einen virtuellen Wert für das Vorkommen eines Bezeichners von $\top$. Wir werden dies benötigen, um Definitionslücken beim Ähnlichkeits- und Unähnlichkeitsmaß zu vermeiden.
Im Falle einer Ontologie wie WordNet werden pro Begriff mehrere Begriffsbezeichner herangezogen. Die Übertragung auf unseren Fall verrechnet für jeden Begriff b genau einen Bezeichner $D(b)$, so dass wir in Gleichung 5.10 mit einfachen Summen auskommen. Wir erhalten ferner für den Informationsgehalt $I(b)$ für $b \in B$ einer gegebenen Ontologie Ω die Darstellung

$$I(b) := -\ln[N'^{-1} \sum_{b_i \leq b} count(D(b))] \tag{5.11}$$

Um die Ähnlichkeit nach Resnik zu berechnen, bestimmt man für ein gegebenes Paar von Begriffen a, b den gemeinsamen Oberbegriff o(a,b), der die Bedingungen

$$a \leq o(a,b), b \leq o(a,b) \tag{5.12}$$

mit

$$\{c_i \in B | a \leq c_i \leq o(a,b), b \leq c_i \leq o(a,b)\} = \emptyset \tag{5.13}$$

erfüllen muss. Da es möglicherweise mehrere solcher Begriffe gibt, deren Gesamtheit wir als Menge $O(a,b)$ notieren können, folgt für die Ähnlichkeit S^0_Ω nach Resnik

$$S^0_\Omega(a,b) = \min_{x \in O(a,b)} I(x) \tag{5.14}$$

.

Diese Berechnung sei anhand eines Beispiels und der Abbildung 5.3 illustriert.

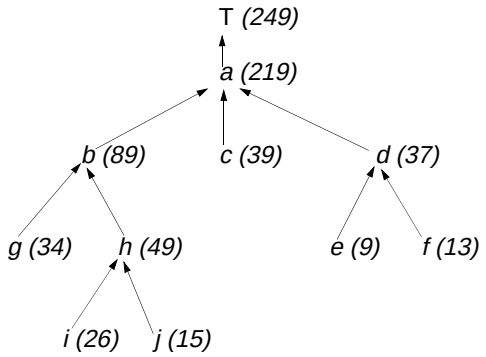

Abbildung 5.3: Aufsummierung von Wortauftrittshäufigkeiten nach Resnik

In unserem Beispiel werden die Begriffe g und j betrachtet. Der gemeinsame Oberbegriff $o(g, j)$ der beiden Begriffe ist b. Die Auftrittshäufigkeit von b, also die Summe der Auftrittshäufigkeiten aller Unterbegriffe von b und von b selbst, beträgt 89. Die Gesamtauftrittshäufigkeit aller Begriffe dieser Ontologie beträgt 249. Somit beträgt die Wahrscheinlichkeit von b als Begriff gerundete 0.36. Die Ähnlichkeit zwischen g und j hat somit den Wert $S^0_\Omega(x, y) = 1.03$.
Das ontologische Vergleichsmaß nach Resnik lässt sich in ein Unähnlichkeitsmaß überführen, wenn der Ansatz der innerhalb der Ontologie vom Speziellen zum Allgemeinen steigenden Wahrscheinlichkeit umgekehrt wird. Daraus erhalten wir als Umformung der Gleichung 5.11 für $b \in B \backslash \top$

$$I'(b) := -\ln[1 - (N'^{-1} \sum_{b_i \leq b} count(D(b)))] \tag{5.15}$$

und als Unähnlichkeitsmaß D^0_Ω für Begriffe gegebene a und b

$$D^0_\Omega(a, b) = \max_{x \in O(a,b)} I'(x). \tag{5.16}$$

Die Minima- und Maximabildung in den Gleichungen 5.14 und 5.16 wird im Falle eines eindeutigen gemeinsamen Oberbegriffes überflüssig, da dann $|O(a,b)| = 1$ gilt.
Mit Unähnlichkeitsmaßen wandelt sich Gleichung 5.7 zu

$$P(b) := \{x \in C | k_{opt} S_v(x,y) \leq T_b\}. \tag{5.17}$$

Dies bedeutet, dass für die Erzeugung eines Vorschlags eine Unähnlichkeitsschranke T_b unterschritten werden muss. Dies führt durch Umformulierung des Algorithmus 2 zu Algorithmus 3, der als Variante von Algorithmus 2 zu betrachten ist.

Data : eine Ontologie $\Omega := (B, \leq, R, \sigma)$, eine optimale Gewichtung $\vec{k}_{opt}$, ein Textkorpus ξ, ein vektorwertiges Unähnlichkeitsmaß S_v, eine reellwertige Schranke T_b pro Begriff, Kandidaten C
Result : Begriffsvorschläge $P(b)$ für alle $b \in B \backslash \{\top\}$
for *alle* $b \in B \backslash \{\top\}$ **do**
  $P(b) \leftarrow \varnothing$
  for *alle* $c \in C$ **do**
    $D_c \leftarrow \vec{k}_{opt} S_v(b,c)$
    if $D_c < T_b$ **then**
      $P(b) \leftarrow P(b) \cup \{c\}$
    end
  end
end

Algorithmus 3 : Unähnlichkeitsbasierte Ontologieanreicherung

Bei Algorithmus 3 kann die Zuordnung $D_c \leftarrow \vec{k}_{opt} S_v(b,c)$ von der ebenfalls möglichen Zuordnung $D_c \leftarrow \vec{k}_{opt} S_v(c,b)$ abweichen. Daher existieren zwei Varianten der unähnlichkeitsbasierten Ontologieanreicherung.

5.4.1.2 Das ontologische Vergleichsmaß nach Li

Li führt in [63] graphentheoretische Maße ein, die allesamt als Ähnlichkeitsmaße definiert sind. Der Graph wird dabei von den Begriffen als Eckenmenge und den Unterbegriffsrelationen $\leq$ als Kantenmenge gebildet. Lis Vorgehensweise basiert auf der Annahme, dass Ähnlichkeit zwischen Begriffen kontextabhängigen und asymmetrischen Effekten unterliegt.

Verbinden wir das graphentheoretische Verständnis ontologischer Ähnlichkeitsmaße mit der Sicht der formalen Begriffsanalyse (Definitionen 2 und 3), so kann eine mögliche Asymmetrie ebenfalls plausibel erscheinen. Es lässt sich mit dem unterschiedlichen Gewinn und Verlust an Information bei Aufwärts- und Abwärtsschritten entlang der Relationen $\leq$ und $\geq$ argumentieren. Schritte vom Unterbegriff zum Oberbegriff sind mit einem Verlust an Merkmalen verbunden, umgekehrt impliziert ein Schritt vom Oberbegriff zu einem Unterbegriff einen Gewinn an Merkmalen. Allerdings bleibt es aufgrund der Vererbung von Merkmalen im ersten Fall (Schritt entlang $\geq$) eindeutiger, wie viele Merkmale erhalten bleiben können: die Mächtigkeit der Merkmalsmenge des Oberbegriffes ist durch die Mächtigkeit der Merkmalsmenge seines Unterbegriffes beschränkt. Beim Hinzutreten neuer Merkmale (Schritt entlang $\leq$) ist eine solche Beschränkung nicht zwingend gegeben. Zudem sind die Merkmale nach Definition 4 bei der Erstellung der Ontologie vom Ersteller und beim Benutzen der Ontologie vom Nutzer gekapselt. Aus dieser Betrachtungsweise erscheint die Annahme von Asymmetrie vorsichtiger als die Annahme von Symmetrie.
Li folgert jedoch, dass Asymmetrieeffekte zu geringe Auswirkungen zeigen, um bei ontologischen Vergleichsmaßen eine definitorische Rolle zu spielen.
Li identifiziert für die ontologiebasierte Ähnlichkeit zweier Begriffe drei Einflussfaktoren.

- die minimale Pfadlänge zwischen den Begriffen im Graphen Γ mit Begriffen (Ecken) und Unterbegriffsrelationen (Kanten)
- die Abstraktionsebene des direkten Oberbegriffs zweier Begriffe
- die aus einem fachspezifischen Textkorpus errechneten Dichte.

Als Abstraktionsebene definieren wir:

Definition 12 (Abstraktionsebene) *Sei $\top$ der abstrakte Wurzelbegriff einer Ontologie. Die Länge des kürzesten Pfades in Γ zwischen $\top$ und einem Begriff b der Ontologie nennen wir Abstraktionsebene von b.*

In der Abbildung 5.4 sind die Abstraktionsebenen für verschiedene Begriffe eingetragen.
Die empirischen Untersuchungen Lis basieren auf Befragungen von Personen, die eine Liste von Begriffspaaren bezüglich ihrer Ähnlichkeit ordnen sollten. Das Maß, welches nach Li am besten dazu in der Lage war, eine den Umfrageergebnissen ähnliche Liste zu erzeugen, wird im Folgenden definiert. Es seien zwei Begriffe a und b gegeben. Sei $h(a, b)$ die Abstraktionsebene des

gemeinsamen Oberbegriffes $o(a, b)$ für a und b, für den $a \leq o(a, b), b \leq o(a, b)$ gilt und für den die Abstraktionsebene aller anderen gemeinsamen Oberbegriffe von a und b kleiner als die Abstraktionsebene von $o(a, b)$ wäre. Sei außerdem $l(a, b)$ der kürzeste Pfad in Γ zwischen a und b. Dann definiert Li die Ähnlichkeitsfunktion, die wir im Folgenden mit S^1_Ω bezeichnen möchten, als

$$S^1_\Omega(a, b) = e^{-\alpha l(a,b)}(e^{\beta h(a,b)} - e^{-\beta h(a,b)})(e^{\beta h(a,b)} + e^{-\beta h(a,b)})^{-1}, \qquad (5.18)$$

wobei α und β die jeweils die Einflüsse der Parameter (Abstraktionsebene und Pfadlänge) gewichten. Dieses Ähnlichkeitsmaß S^1_Ω ist durch einfache Subtraktion für unsere Zwecke in ein Unähnlichkeitsmaß D^1_Ω überführbar:

$$D^1_\Omega(a, b) = 1 - S^1_\Omega(a, b) \qquad (5.19)$$

Zur Veranschaulichung verweisen wir auf Abbildung 5.4. Als Länge $l(j, d)$ des kürzesten Pfades zwischen den Begriffen j und d erhalten wir $l(j, d) = 4$. Der gemeinsame Oberbegriff von j und d ist hier a mit der Abstraktionsebene $h(j, d) = 1$. Für die von Li vorgeschlagenen Parameter $\alpha = 0.2$ und $\beta = 0.6$ berechnen wir $S^1_\Omega(j, d) = 0.24$ beziehungsweise $D^1_\Omega = 0.76$.

5.4.1.3 Das asymmetrische Vergleichsmaß

Die ursprüngliche Konzeption des asymmetrischen ontologischen Vergleichsmaßes erfolgte zur Visualisierung von Ausschnitten einer Ontologie und als Hilfswerkzeug beim Vergleich von verschlagworteten Textdokumenten [27]. Unsere Untersuchungen werden die Annahme Lis, dass Asymmetrie bei der Definition von ontologischen Vergleichsmaßen vernachlässigt werden kann, für Fragestellungen der Ontologieanreicherung nochmals einer Prüfung unterziehen.
In unserem Beitrag zu ontologischen Vergleichsmaßen, dem asymmetrischen Vergleichsmaß, übernehmen wir prinzipielle psycholinguistische Annahmen für die semantische Unähnlichkeit zweier Begriffe einer Ontologie. Eine längere minimale Pfadlänge zwischen zwei Begriffen erhöht die Unähnlichkeit, da die beiden Begriffe mit größerer Wahrscheinlichkeit über einen relativ abstrakten Oberbegriff verfügen, der einen geringen Informationsgehalt hat. An dieser Stelle kommt somit die Grundannahme Resniks zum Informationsgehalt von Begriffen zum Tragen. Auch die formale Begriffsanalyse zeigt für abstrakte Begriffe einen geringe Begriffsintension (vergleiche Definition 2), was nach [62] zur Berechnung von Ähnlichkeiten und Unähnlichkeiten

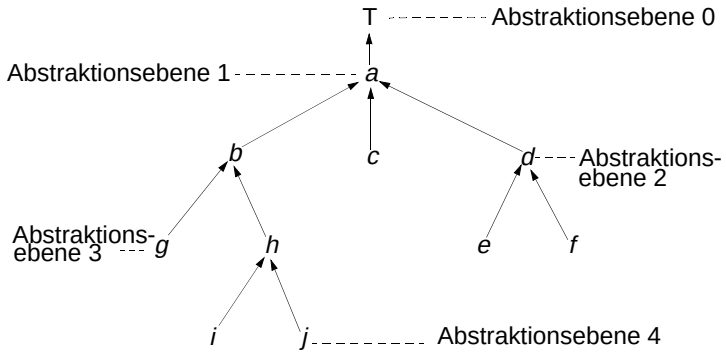

Abbildung 5.4: Das ontologische Vergleichsmaß nach Li

herangezogen werden kann.
Schritte von oben nach unten in der Hierarchie führen, wie im letzten Abschnitt begründet, zu einer größeren Unähnlichkeit als umgekehrt, genauer gesagt schließt beispielsweise der Begriff *Ente* den Informationsgehalt von *Vogel* ein, umgekehrt jedoch nicht.

Für eine im asymmetrischen Vergleichsmaß singuläre weitere Grundannahme definieren wir begriffliche Geschwister.

Definition 13 (Geschwister) *Sei* $\Omega := (B, \leq, R, \sigma)$ *eine gegebene Ontologie und* $b, c \in B$ *Begriffe. Wenn* $b \leq c$*, dann nennen wir die Menge* $\{b_i \in B | b_i \leq c, \{b_i^{'} | b_i < b_i^{'} < c\} = \emptyset\}\} \cup \{b\}$ *die Menge der Geschwister von* b*.*

Geschwister erhöhen die Unähnlichkeit. Das heißt, je mehr Unterbegriffe ein Begriff b besitzt, desto höher ist die Unähnlichkeit zwischen dessen Unterbegriffen, aber auch zwischen b und seinen Unterbegriffen. Wir setzten dies voraus, da wir annehmen, dass das Auftreten vieler direkter Unterbegriffe zu einem Begriff auf fehlende Abstraktionsebenen in der Hierarchie hindeutet. Zur systematischen Redefinition von Ontologien mit zusätzlichen Abstraktionsebenen sei an dieser Stelle auf Arbeiten Guarinos und Weltys [103]

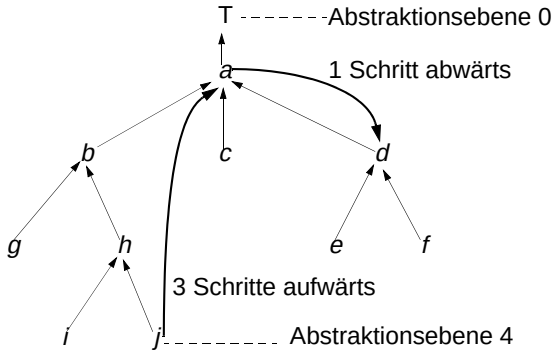

Abbildung 5.5: Aufwärts- und Abwärtsschritte entlang der Unterbegriffsrelation

verwiesen. Wir stützen unsere Annahme möglicherweise fehlender Abstraktionsebenen auf diese Arbeit. Guarino und Welty arbeiten zwar mit einer Logik von Metaeigenschaften der Begriffe, die über die rein graphentheoretische Betrachtung unserer Geschwisterdefinition hinausgeht, kommen jedoch ebenfalls zu dem Schluss, dass schon bei einer geringen Anzahl von Geschwistern Abstraktionsebenen fehlen können und dass sich diese Situation bei vielen Geschwistern tendenziell verschärft.
Für isomorphe Pfade, das heißt für die gleiche Abfolge von Auf- und Abwärtsschritten, ist die Unähnlichkeit zwischen abstrakten Begriffen höher als zwischen speziellen Begriffen. Das bedeutet, dass das asymmetrische Vergleichsmaß die Zusammenhänge auf abstraktem Niveau als vage im Vergleich zu denen auf einer konkreteren Ebene beurteilt. Diesen Zusammenhang kann man mit dem geringeren Informationsgehalt auf abstrakten Ebenen begründen. Die Argumentation stützt sich somit wieder auf die von Resnik eingeführten Instrumente, ohne dass wir hier direkt den Informationsgehalt von Begriffen berechnen.
Es seien wiederum zwei Begriffe a, b einer bestehenden Ontologie gegeben. Die einzelnen in die Berechnung des asymmetrischen Vergleichsmaßes S^2_Ω

eingehenden Größen sind der minimale Pfadlänge zwischen a und b in G, dessen Schritte aufwärts in der Ober-Unterbegriffshierarchie mit κ, die Schritte abwärts mit λ notiert werden, die durchschnittliche Abstraktionsebene α von a und b und schließlich γ, die durchschnittliche Anzahl von Geschwistern von a und b. Innerhalb der Berechnungen werden wir zum Schließen von Definitionslücken die Schrittanzahlen κ und λ um 1 erhöhen und definieren nun als asymmetrisches Vergleichsmaß S^2_Ω:

$$S^2_\Omega := (\gamma^{-1}\alpha((\kappa+1)^{-1} + (2\lambda+2)^{-1}))^2 \tag{5.20}$$

Dieses Maß berücksichtigt die Asymmetrie, weil die Abwärtsschritte λ doppelt so stark eingehen wie die Aufwärtsschritte κ und somit die Ähnlichkeit stärker mindern. Spezialisierende Schritte (Abwärtsschritte) fügen aus der Sicht der formalen Begriffsanalyse Merkmale zum Oberbegriff hinzu. Die Vereinbarkeit dieser hinzugewonnenen Merkmale mit dem Ausgangsbegriff eines Pfades (entlang der Begriffe und Ober-Unterbegriffsrelationen) ist nicht zwingend gegeben. Umgekehrt sind die bei Aufwärtsbewegungen erhaltenen Merkmale eines Oberbegriffes stets auch Merkmale des Unterbegriffes. Die Wertungen durch κ und λ tragen diesem Zusammenhang Rechnung.
Eine hohe Geschwisteranzahl γ führt zu einer vergleichsweise niedrigeren Ähnlichkeit. Eine hohe Abstraktion geht mit einem niedrigen α einher. Das heißt, die Nähe zum abstrakten Wurzelbegriff $\top$, wirkt ebenfalls mindernd auf die Ähnlichkeit.
Da wir nicht von einem festen Intervall ausgehen können, in dem sich S^2_Ω bewegt, andererseits aber $S^2_\Omega > 0$ gilt, formen wir das asymmetrische Vergleichsmaß durch Kehrwertbildung in ein Unähnlichkeitsmaß D^2_Ω um:

$$D^2_\Omega := \frac{1}{S^2_\Omega}. \tag{5.21}$$

Die Definitionslücke für D^2_Ω unter $S^2_\Omega = 0$ wird in unserer Anwendung keine Rolle spielen, denn zu $\top$ liegt nie ein Begriffsvektor vor. $\top$ geht nicht in das Minimierungsproblem 5.4 ein. Die Tatsache, dass die Unähnlichkeit D^2_Ω eines Begriffes zu sich selbst größer als 0 sein und unterschiedliche Werte für den Vergleich von Begriffen mit sich selbst entstehen können, muss in unserem Zusammenhang keinen Nachteil darstellen. Zum einen ist letzteres auch schon bei der Ähnlichkeitsformulierung nach Resnik der Fall, zum anderen wird als Eingangskriterium (T_b nach Gleichung 5.8) für einen neuen Begriffsvorschlag stets die Ähnlichkeit oder Unähnlichkeit zu benachbarten Begriffen herangezogen. In die Berechnung von T geht der Ähnlichkeits- oder Unähnlichkeitswert eines Begriffes zu sich selbst nicht ein.

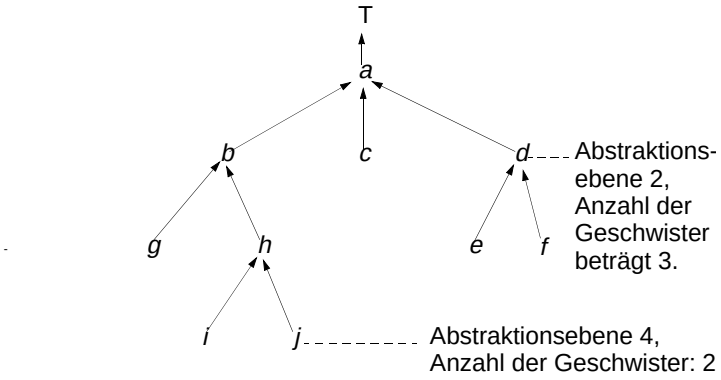

Abbildung 5.6: Begriffliche Geschwister

Es bleibt festzuhalten, dass gerade bei der Parametrisierung des Unterschieds zwischen Schritten entlang $\leq$ und Schritten entlang $\geq$ auch noch zahlreiche andere Werte als die Verdopplung von λ denkbar sind. Im Rahmen der vorliegenden Arbeit fand eine Beschränkung auf obige Asymmetrieheuristik statt.
Wiederum geben wir ein Beispiel für die Berechnung des vorgestellten Maßes und verweisen dazu auf Abbildung 5.6.

Wir vergleichen d und j. Der Begriff j befindet sich in Abbildung 5.6 auf Abstraktionsebene 4 der Ontologie, der Begriff d auf Abstraktionsebene 2. Daraus ergibt sich $\alpha = 3$. Der Begriff j besitzt nach Definition 13 zwei Geschwister, d besitzt drei Geschwister. Man erhält als durchschnittliche Anzahl der Geschwister $\gamma = 2.5$. Von d nach j durchläuft man auf dem kürzesten Pfad $\kappa = 3$ Schritte aufwärts und $\lambda = 1$ Schritt abwärts im anhand der Ontologie definierten Graphen G. Wir erhalten mit diesen Werten $D^2_\Omega = 2.8$ (gerundet), beziehungsweise $S^2_\Omega = 0.36$.
Im Vergleich zu den Maßen nach Resnik und Li gilt im Allgemeinen nicht $S^2_\Omega(a, b) = S^2_\Omega(b, a)$. Dies wurde über die Unterscheidung von abstrahierenden und konkretisierenden Schritten innerhalb der Ober-Unterbegriffsstruktur

erreicht. Ein weiterer originärer Beitrag ist die Berücksichtigung der Geschwisterknoten. Ziel hierbei ist es, unzureichend ausmodellierte Abstraktionsebenen der Ontologie erfassen zu können. Schließlich stellt die Behandlung der Abstraktionsebene zu vergleichender Begriffe durch das asymmetrische Vergleichsmaß ein Novum dar, da hier im Gegensatz zu Resnik und Li das durchschnittliche Abstraktionsniveau und nicht das Abstraktionsniveau eines gemeinsamen Oberbegriffes herangezogen wird.

5.4.2 Vektorwertige Vergleichsmaße

In Definition 10 wurde eine Festlegung vektorwertiger Vergleichsmaße getroffen. Der Zweck dieser vektorwertigen Vergleichsmaße liegt darin, Ähnlichkeiten und Unähnlichkeiten, die aufgrund des Vergleiches der in Vektoren festgehaltenen Kollokationseigenschaften von Begriffen berechnet werden, zunächst als Vektor zu erfassen. Der eigentliche Anreicherungsansatz bildet ein Skalarprodukt aus diesem (aus einem Vergleich resultierenden) Vektor und einem durch als Lösung des Minimierungsproblems 5.4 gewonnenen Vektor. Die Lösung des Minimierungsproblems 5.4 liefert auch die Anpassung der resultierenden Ähnlichkeiten oder Unähnlichkeiten zwischen Paaren von Begriffsbezeichnern an die Gestalt der Ontologie. Einzelne Koordinaten der Begriffsvektoren und somit der Einfluss einzelner Kollokatoren werden für alle Resultate vektorwertiger Vergleiche von Begriffspaaren stärker oder schwächer gewichtet.
Im Folgenden werden zwei Varianten vektorwertiger Vergleichsmaße eingeführt. Es handelt sich dabei um eine Übertragung der reellwertigen Vergleichsmaße jeweils zweier Vektoren nach Lillian Lee [59]. In ihrer Arbeit zur Vorhersage von Wörtern in amerikanisch-englischen Sätzen führt sie ausführliche Untersuchungen der Auswirkungen verschiedener formalisierter Vergleiche von Vektoren auf die Vorverarbeitung der Daten durch. Diese Analysen und ihre empirischen Untersuchungen weisen einen starken Bezug zu Clusterverfahren auf. Teile der Vergleichsmaße Lees wurden von Mädche, Pekar und Staab für clusterbasierte automatische Ontologiekonstruktionen verwendet [68]. Sowohl in den empirischen Arbeiten Lees als auch in der empirischen Arbeit Mädches, Pekars und Staabs schneiden Vergleichsmaße, die nach dem Ähnlichskeitsverständnis des Jaccard-Maßes und solche, die nach dem Unähnlichkeitsverständnis der Kullback-Leibler-Divergenz und des aus ihr resultierenden Schiefmaßes formuliert wurden, signifikant am besten ab. Die im folgenden Abschnitt zu treffende Neuformulierung des Jaccardmaßes und des Schiefmaßes als vektorwertige Vergleichsmaße stellt einen originären eigenen Beitrag der vorliegenden Arbeit dar. Neben der Schreibweise für be-

liebige Vektoren gleicher Länge werden die Notationen mittels der Repräsentationsmatrix aus Abschnitt 5.3.1 fortgeführt.

5.4.2.1 Ein vektorwertiges Vergleichsmaß auf Basis des Jaccardmaßes

Dem Jaccardmaß liegt die Vorstellung zugrunde, dass der Anteil der Schnittmenge von erfüllten Attributen des Sprachmodells an der Vereinigungsmenge der Attribute aus dem Sprachmodell eine Ähnlichkeitsaussage liefert. Seien dazu $A, B \subseteq \mathcal{A}$, das heißt, Teilmengen der Attributmenge. Dann ist das Jaccardmaß $J(A, B)$ als Koeffizient

$$J(A, B) := \frac{|A \cap B|}{|A \cup B|} \tag{5.22}$$

definiert.
Im Falle der in 5.3.1 eingeführten Repräsentationsmatrix, deren Zeilen als Vektoren bezüglich $\mathcal{A}$ aufzufassen sind, ist eine Anpassung der Mengenoperation aus der vorherigen Gleichung 5.22 notwendig.
Prinzipiell bestehen zwei Anpassungsmöglichkeiten. Wenn die Einträge der auf einem Sprachmodell $\mathcal{A}$ beruhenden Vektoren $\vec{a}$ und $\vec{b}$ jeweils entweder den Wert 0 oder den Wert 1 annehmen, so können $\vec{a}$ und $\vec{b}$ zu Mengen $M(a), M(b) \subseteq \mathcal{A}$ umgeformt werden. Dies geschieht durch die Aufnahme des i-ten Attributs $a_i \in \mathcal{A}$, falls der Vektor in der i-ten Koordinate den Wert 1 annimmt. Bei dieser Vorgehensweise muss im Falle des kollokationsbasierten Sprachmodells, welches hier genutzt wird, allerdings ein Mechanismus vorliegen, der die Übersetzung der Korpusdaten in Vektoreinträge aus der Menge $\{0, 1\}$ vornimmt. Denkbar ist hier, in Anlehnung an beispielsweise [42], ein Eintrag 1 im Falle signifikanter Vorkommnisse einer Kollokation. Dies erfordert eine zusätzliche Festlegung in Form eines Schwellwertes, dessen Überschreitung die Signifikanz anzeigt.
Eine Alternative besteht darin, das j-fache Vorkommen eines Attributes $a \in \mathcal{A}$ als j verschiedene Attribute zu interpretieren. Diese werden allerdings in unserer Interpretation in a gebündelt.
Seien $|\mathcal{A}| = n$ und $\vec{a}$ und $\vec{b}$ Vektoren aus $\mathbb{R}^n$, die nach dem Sprachmodell gebildet wurden. Dann kann besagte Interpretation zu folgender Umformulierung des Jaccard-Maßes für die Vektoren genutzt werden:

$$J(\vec{a}, \vec{b}) := \sum_{i=1}^{n} \frac{\min(a_i, b_i)}{\sum_{l=1}^{n} \max(a_l, b_l)}, \tag{5.23}$$

wobei der min-Operator die Mächtigkeiten der Schnittmengen ermittelt. Der max-Operator und die Summierung im Nenner normieren die Summanden der äußeren Summe über die Mächtigkeiten der Vereinigungsmengen. Ein Ziel der experimentellen Untersuchungen im folgenden Kapitel wird es sein, diese neuartige Formulierung eines Jaccard-Maßes Tests zu unterziehen. Der Schwerpunkt dieser Tests wird dabei auf der Gewichtung jedes einzelnen Summanden zur Lösung des Minimierungsproblems 5.4 liegen. Wir ziehen dazu für die Summanden aus Gleichung 5.23 folgendes vektorwertiges Vergleichsmaß nach Definition 10 heran. Für die i-te Komponente $\vec{v}_i$ des resultierenden Vektors $\vec{v}$ gelte:

$$\vec{v}_i := \frac{\min(a_i, b_i)}{\sum_{j=1}^{n} \max(a_j, b_j)} \tag{5.24}$$

Mit Hilfe der Repräsentationsmatrix $M := m_{ij}$ ergibt sich eine vektorwertige Ähnlichkeitsbestimmung zweier Begriffe in einer gegebenen Ontologie durch den Vergleich der Einträge der i-ten und der j-ten Zeile. Die $i-$te Komponente des dazu gehörigen Vektors $\vec{v}(kj)$ wird durch

$$\vec{v}(kj)_i := \frac{\min(m_{ki}, m_{ji})}{\sum_{l=1}^{n} \max(m_{kl}, m_{jl})} \tag{5.25}$$

bestimmt und wird somit als Teil des Minimierungsproblems 5.4 überführbar. Die ontologischen Vergleichsmaße sind in diesem Fall jeweils in der **Ähnlichkeitsvariante** zu verwenden.

5.4.2.2 Vektorwertige Vergleichsmaße auf Grundlage der Kullback-Leibler-Divergenz

In den Untersuchungen Lees wird die besondere Eignung von Maßen nachgewiesen, die auf der Kullback-Leibler-Divergenz beruhen. Diese bietet eine Methode, die Unähnlichkeit zwischen zwei Wahrscheinlichkeitsverteilungen zu bestimmen. Da die Kullback-Leibler-Divergenz in ihrer ursprünglichen Formulierung $D(\vec{a}||\vec{b})$ mit n-dimensionalen Vektoren $\vec{a}, \vec{b}$ und Vektorkoordinaten a_i, für $\vec{a}$ und b_i für $\vec{b}$, $1 \leq i \leq n$ durch die Gleichung

$$D(\vec{a}||\vec{b}) := \sum_{i=1}^{n} \frac{a_i}{\sum_{j=1}^{n} a_j} (\ln \frac{a_i / \sum_{l=1}^{n} a_l}{b_i / \sum_{p=1}^{n} b_p}) \tag{5.26}$$

gegeben ist, tauchen bei Sprachdaten, die häufige $a_i = 0$ oder $b_i = 0$ aufweisen, Definitionslücken auf. Das Problem der durch den Logarithmus aufgeworfenen Definitionslücken, auch als Problem dünn besetzter Daten

bei Lee behandelt, gilt somit auch für einen hier neu zu formulierenden vektorwertigen Vergleich zweier Zeilen $\vec{z_i}$ und $\vec{z_k}$ der Repräsentationsmatrix mit Koordinaten $D(z_i||z_k)_l$:

$$D(\vec{z_i}||\vec{z_k})_l := \frac{m_{il}}{\sum_{j=1}^{n} m_{ij}}(\ln \frac{m_{il}/\sum_{j=1}^{n} m_{ij}}{m_{kl}/\sum_{j=1}^{n} m_{kj}}) \tag{5.27}$$

Lee führt das auf der Kullback-Leibler-Divergenz basierende Schiefmaß D_α als

$$D_\alpha(\vec{a}||\vec{b}) := D(\vec{b}||\alpha\vec{a} + (1-\alpha)\vec{b}) \tag{5.28}$$

und erzielt damit für das Wortcluster und -vorhersageproblem im Vergleich zu den anderen Ähnlichkeits- und Unähnlichkeitsheuristiken die besten Ergebnisse. Der Parameter α liegt in den Experimenten von Lee optimalerweise bei 0.99. Für unsere Zwecke definieren wir ausgehend von Gleichung 5.28 ein vektorwertiges Schiefmaß als Modifikation von Gleichung 5.28. Der Vergleich der i-ten Zeile $\vec{z_i}$ mit der k-ten Zeile $\vec{z_k}$ der Repräsentationsmatrix erfolgt dabei nun durch ein

$$(D_\alpha)(\vec{z_i}||\vec{z_k})_l := \frac{m_{kl}}{\sum_{j=1}^{n} m_{kj}}(\ln \frac{m_{kl}/\sum_{j=1}^{n} m_{kj}}{(\alpha m_{il})(1-\alpha)m_{kl}/\sum_{j=1}^{n}(\alpha m_{ij} + (1-\alpha)m_{kj})}) \tag{5.29}$$

Diese Anwendung des Schiefmaßes enthält allerdings weiterhin Definitionslücken, und zwar für die Fälle:

$$\exists k, 1 \leq k \leq n : m_{kl} = 0 \tag{5.30}$$

Um die Anzahl dieser Fälle zu reduzieren, transformieren wir zunächst Gleichung 5.28, so dass jeweils beide Komponenten eines Vektors an Einfluss auf das Unähnlichkeitsmaß gewinnen und wir ein Maß D_α^g erhalten:

$$D_\alpha^g(\vec{a}||\vec{b}) := D(\alpha\vec{b} + (1-\alpha)\vec{a}||\alpha\vec{a} + (1-\alpha)\vec{b}) \tag{5.31}$$

Die Gleichung 5.29 nimmt dadurch die Form

$$D_\alpha^g(\vec{z_i}||\vec{z_k})_l :=$$

$$\frac{\alpha m_{kl} + (1-\alpha)m_{il}}{\sum_{j=1}^{n}(\alpha m_{kj} + (1-\alpha)m_{ij})}(\ln \frac{\alpha m_{kl} + (1-\alpha)m_{il}/\sum_{j=1}^{n}(\alpha m_{kj} + (1-\alpha)m_{ij})}{(\alpha m_{il})(1-\alpha)m_{kl}/\sum_{j=1}^{n}(\alpha m_{ij} + (1-\alpha)m_{kj})}) \tag{5.32}$$

an.

Damit sind auch für den Unähnlichkeitsfall neuartige vektorwertige Formulierungen der Vergleichsmaße erreicht; wir nennen D_α das vektorwertige Schiefmaß und D_α^g das geglättete Schiefmaß.

5.4.2.3 Beispiel

Dieser Abschnitt zeigt ein Beispiel der ähnlichkeitsbasierten Ontologieanreicherung.
Gegeben ist die Ontologie mit den Begriffen $\top$, *Krankheit ODER Symptom*, *Darminfektion*, *Durchfall*, *Masern*, *Röteln*, *blutiger Durchfall*, *wässriger Durchfall* und den Oberbegriffsbeziehungen *Krankheit ODER Symptom* $\geq$ *Durchfall*, *Krankheit ODER Symptom* $\geq$ *Darminfektion*, *Krankheit ODER Symptom* $\geq$ *Masern*, *Krankheit ODER Symptom* $\geq$ *Röteln*, *Durchfall* $\geq$ *blutiger Durchfall*, *Durchfall* $\geq$ *wässriger Durchfall*.
Aus einem Textkorpus, der aus den Dokumenten der jeweils zehn am höchsten platzierten Google-Suchtreffer zu jedem der Bezeichner bestand, wurden die Kollokationsinformationen gewonnen. Die Kollokationseigenschaft, die Berücksichtigung fand, war das Vorkommen eines Bezeichners mit den Wörtern im maximalen Abstand $\delta_W = 5$ im Textkorpus. Die Kollokationsinformationen zu *Krankheit* und *Symptom* wurden dabei aufaddiert.
Als vektorwertiges Vergleichsmaß wurde das modifizierte Jaccardmaß, als ontologisches Vergleichsmaß das asymmetrische Vergleichsmaß gewählt.
Es sollten aus Gründen der Übersichtlichkeit höchstens 30 Begriffsvorschläge für die Ontologie gewählt werden, was mit einer Schranke von $T(b) = 0.21$ für alle Begriffe b aus der Beispielontologie erreicht werden konnte. Tabelle 5.2 gibt einen Überblick über die Ergebnisse.

Tabelle 5.2: Ontologieanreicherungsbeispiel

Begriff	*Vorschlag*
Krankheit ODER Symptom	*Ausbruch,Infektion,motorisch,bakteriell*
Darminfektion	*Infektionen, bakteriell,besteht,beschreibt, bietet, Enzephalitis,Komplikationen*
Durchfall	*Adlerfarn, motorisch,nervös*
Masern	*bakteriell, besteht, beschreibt, bietet, Enzephalitis, Informationshotline, Komplikationen, medifon, motorisch*
Röteln	*nervös*
blutiger Durchfall	*Ansteckung,Infektion,Viren*
wässriger Durchfall	*nervös*

Zur Bewertung des Ergebnisses ist die Expertise eines Mediziners notwendig. Wir geben daher nur einen kurzen Kommentar ab. Die Vorschläge von *Infektion*, *Enzephalitis*, *bakteriell*, *Viren*, *Ausbruch*, *Ansteckung* und

Tabelle 5.3: Maßkombinationen und Vergleichsarten

ontologisches Vergleichsmaß	*vektorwertiges Vergleichsmaß*	Vergleich
Jaccard	Resnik	Ähnlichkeit
Jaccard	Li	Ähnlichkeit
Jaccard	asymmetrisches Vergleichsmaß	Ähnlichkeit
Schiefmaß	Resnik	Unähnlichkeit
Schiefmaß	Li	Unähnlichkeit
Schiefmaß	asymmetrisches Vergleichsmaß	Unähnlichkeit
geglättetes Schiefmaß	Resnik	Unähnlichkeit
geglättetes Schiefmaß	Li	Unähnlichkeit
geglättetes Schiefmaß	asymmetrisches Vergleichsmaß	Unähnlichkeit

Komplikationen sind eher positiv und als dem Thema entsprechend zu werten, die Vorschläge von *Adlerfarn*, *medifon*, *Informationshotline* sind korpusspezifisch, da zahlreiche medizinische Informationssdienste als Suchtreffer vorlagen. Als weniger brauchbar erweisen sich die Vorschläge der Tätigkeitswörter (wie *besteht*, *beschreibt* et cetera), und bei der Interpretation der Adjektive *nervös* und *motorisch* besteht in der Bewertung Spielraum.
Insgesamt ergibt sich das Bild, dass die ähnlichkeitsbasierte Ontologieanreicherung in diesem Fall als Hilfe zur systematischen Erweiterung der Ontologie in Frage kommt.

5.4.2.4 Mögliche Kombinationen aus ontologischen und vektorwertigen Vergleichsmaßen

Wir geben als Abschluss dieses Kapitels nochmals einen tabellarischen Überblick über die möglichen Kombinationen aus ontologischen und vektorwertigen Vergleichsmaßen, die in das Minimierungsproblem 5.4 eingehen können. Diese Kombinationen sollen im nächsten Kapitel Gegenstand ausführlicher empirischer Untersuchungen sein. Die Tabelle 5.3 zeigt, welche Form eines Vergleiches, Ähnlichkeit oder Unähnlichkeit, für die jeweilige Kombination von vektorwertigen und ontologischen Maßen per Konstruktion eingesetzt werden muss. Dazu kann auf die jeweils zweifache Formulierung der ontologischen Vergleichsmaße zurückgegriffen werden. Es bleibt festzuhalten, dass bei der Anreicherung für die Unähnlichkeitsmaße eine reellwertige Schranke unterschritten werden muss, um einen Kandidaten zum Vorschlag werden zu lassen. Dies ist somit bei den Schiefmaßen nach Gleichungen 5.29 und 5.32, der Fall. Analog dazu müssen bei allen Anreicherungsverfahren, die mit Ähnlichkeitsmaßen operieren, reellwertige Schranken überschritten wer-

den, um eine Anreicherung zu erreichen. Dies tritt bei den Ähnlichkeitsmaßen ein, die auf dem für Mengen mit Wiederholungen modifizierten Jaccardmaß basieren. Die Kombination aus dem asymmetrischen Vergleichsmaß und dem modifizierten Jaccardmaß liefert keinen eindeutigen Hinweis zur Bestimmung der im Falle der Anreicherung zu überschreitenden reellwertigen Schranke, weil bei der Bildung einer Schranke mittels der durchschnittlichen Ähnlichkeit eines benachbarten Begriffspaares aus der Ontologie prinzipiell zwei Werte pro Begriffspaar zur Wahl stehen. Da das asymmetrische ontologische Vergleichsmaß ohnehin für die asymmetrischen vektorwertigen Vergleichsmaße konzipiert wurde, findet die Kombination aus S^2_Ω und dem vektorwertigen Jaccardmaß im Laufe der Untersuchungen keine Berücksichtigung mehr.
Im folgenden Abschnitt werden Evaluationsmöglichkeiten für Ontologieanreicherungsverfahren entwickelt. Damit wird sowohl der Vergleich der in der Tabelle 5.3 aufgeführten Kombinationen und der dadurch entstehenden Anreicherungsverfahren untereinander, als auch der Vergleich zu anderen Anreicherungsverfahren möglich.

5.4.3 Automatische Evaluation

Der folgende Abschnitt stellt Evaluationstechniken für Ontologieanreicherungsverfahren im Allgemeinen dar.
Bei der Definition von Gütemaßen für Ontologieanreicherungsverfahren ist es notwendig, dass all diese Ansätze ohne die Befragung von Nutzern operieren. Dazu seien zwei hauptsächliche Begründungen aufgeführt.
Zum einen ist in einer konzeptionellen Phase, bei der ein Anreicherungsansatz genauer spezifiziert wird und beispielsweise auf eine wie in Tabelle 5.3 angezeigten Parametrisierung hin untersucht wird, der schiere Umfang an Befragungen der am Projekt Beteiligten nicht zu rechtfertigen.
Des Weiteren sind, wie schon in Kapitel 3 erläutert, Fachgebietsexperten für die Ontologieerstellung an sich schwer zu gewinnen. Dieser Umstand verschärft die Schwierigkeit, eine geeignete Expertengruppe zur sukzessiven Evaluation einer Ontologieerstellung zu bewegen, denn die Expertengruppe, welche evaluiert, muss ebenso versiert sein wie die eigentlichen Ersteller.
Für eine gegebene Ontologie $\Omega := \{B, \leq, R, \sigma\}$ verallgemeinern wir die Idee aus [68]. Allerdings entfernen wir zufällig ganze Begriffe aus der gegebenen Ontologie, während Mädche et. al [68] aufgrund der in seinen Arbeiten gegebenen Thesauri von mehrfach vorhandenen Begriffsbezeichnern ausgehen. Die Begriffe können bei Mädche et. al erhalten werden und nur einzelne natürlichsprachliche Bezeichner werden entfernt. Diese Vorgehensweise ist

nicht auf unser Szenario übertragbar. In den Situationen, wo nur genau ein Begriffsbezeichner vorliegt, fehlte es den Anreicherungsverfahren aus Algorithmus 1, Algorithmus 2 und Algorithmus 3 an Eingangsdaten, wenn der Begriffsbezeichner fehlen würde.
Wir sammeln die zufällig entfernten Bezeichner in einer Menge C. Diese dient als Kandidatenmenge, die unabhängig von weiteren Bezeichnern aus dem der Anreicherung zugrunde liegenden Korpus existiert. Die Vorgehensweise mit einer so entstehenden Kandidatenmenge C kann nur dann verfolgt werden, wenn

$$\Omega^{'} := \{B \setminus C, \leq, R', \sigma'\}, \tag{5.33}$$

wieder eine Ontologie ist, wobei R' und σ' Einschränkungen der ursprünglichen Relationen auf $B \setminus C$ darstellen können. Das Anreicherungsverfahren, welches im Rahmen dieser Arbeit entwickelt wird, benutzt als Eingangsdaten allerdings nur $B \setminus C$ und $\leq$, daher können wir auch für Evaluationszwecke $R' = \sigma' = \emptyset$ annehmen.
Solange wir $\top \notin C$ garantieren, gilt dies aber durch die Transitivität der Unterbegriffsrelation. Die Transitivität nimmt bei dem Evaluierungsverfahren eine entscheidende Rolle ein.
Sei nun $d_\Omega(b_1, b_2)$ für $b_1, b_2 \in B$ die Länge des kürzesten Pfades in dem Graphen, der entsteht, wenn wir die Begriffe der gegebenen Ontologie als Ecken und die Unterbegriffsrelationen als Kanten auffassen. Zusätzlich sollen in diesem Graphen analog zu unserer bildlichen Darstellung der Ontologien aus Kapitel 2 diejenigen Kanten, die aufgrund der Transitivität der Unterbegriffsrelation gelten, entfernt werden. Wir definieren unser Maß 1-Kanten-Recall für ein gegebenes $c \in B \setminus C$ als

$$\frac{|P(c) \cap \{b \in B | d_\Omega(c, b) = 1\}|}{|\{b \in B | d_\Omega(c, b) = 1\}|} \tag{5.34}$$

und allgemeiner den **n-Kanten-Recall** für ein gegebenes $c \in B$

$$\frac{|P(c) \cap \{b \in B | d_\Omega(c, b) \leq n\}|}{|\{b \in B | d_\Omega(c, b) \leq n\}|} \tag{5.35}$$

Wenn 5.34 und 5.35 hoch ausfallen, so fällt die Ontologieanreicherung in dem Sinne gut aus, dass eine ursprünglich vorhandene Ontologie korrekt ergänzt wird. Der 1-Kanten-Recall misst dabei die Wiederherstellung von direkten Ober- und Unterbegriffsbezügen, der 2-Kanten-Recall misst darüber hinaus die Wiederherstellung von Geschwisterbezügen und indirekteren Ober- und Unterbegriffsbezügen.

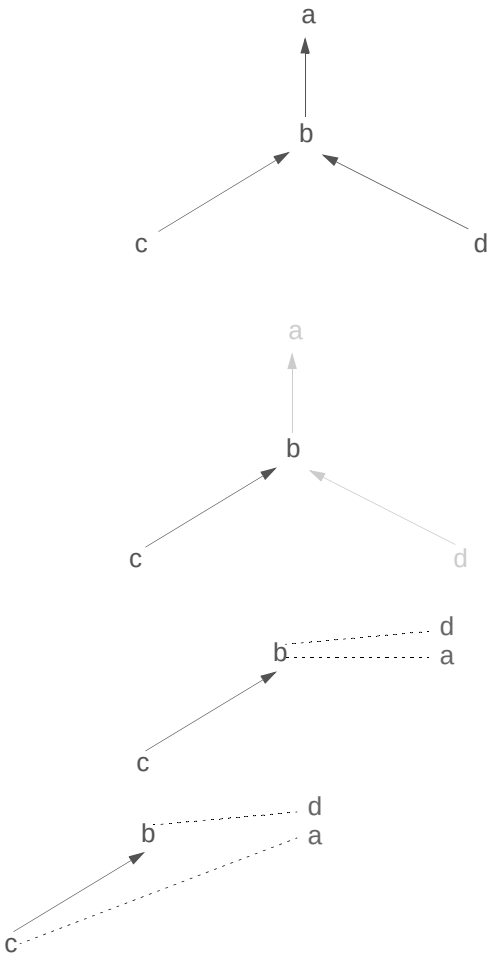

Abbildung 5.7: Recall-Beispiel

Abbildung 5.7 zeigt ein Beispiel zu den n-Kanten-Recallmaßen. Aus einer gegebenen Ontologie (oben) wurden die Begriffe d und a entfernt und wir erhalten somit Kandidaten $C = \{a, d\}$, hier hellgrau gefärbt. Die verbleibenden in der Abbildung 5.7 dargestellten Strukturen zeigen den Anreicherungsprozess. Vorschläge sind wiederum durch eine gestrichelte Linie zwischen Begriff und Kandidat eingetragen.
Im ersten Fall der beiden Anreicherungen betragen der 1-Kanten-Recall und der 2-Kanten-Recall jeweils 100%. Im zweiten Fall (unten) wird ein 1-Kanten-Recall von 100% und ein 2-Kanten-Recall von 50% erreicht.
Durch Variation von $|C|$ kann bestimmt werden, ob ein Anreicherungsverfahren für mehr (kleineres $|C|$) oder weniger (größeres $|C|$) vollständige Ontologien wirkt.
Um die Aussagekraft der Recallmaße auch mit der allgemeinen Neigung des Anreicherungsverfahrens zu Fehlklassifikationen ergänzen zu können, definieren wir das Maß

$$\frac{|P(c) \cap \{b \in B | d_\Omega(c, b) > n\}|}{|\{b \in B | d_\Omega(c, b) > n\}|} \tag{5.36}$$

Nimmt diese so genannte **n-Kanten-Fehlklassifikation** niedrige Werte an, so fallen die Anreicherungen relativ präzise aus.
Eine weiterer Hinweis auf automatisch evaluierbare Formen der Fehlklassifikation kann über die Anreicherungsneigung gewonnen werden.
Von **innerontologischer Anreicherungsneigung**

$$\frac{|P(b)|}{|C|} \tag{5.37}$$

für einen Begriff b sprechen wir, wenn die hier verwendete Kandidatenmenge C thematisch als zur Ontologie gehörig betrachtet werden. Ist dies nicht der Fall, so sprechen wir von 5.37 als **fremdontologischer Anreicherungsneigung**. Letztere ist beispielsweise mit Kandidatenmengen aus thematisch weiter entfernten Bereichen einer Ontologie herstellbar. Voraussetzung hierfür ist, dass das Ontologieanreicherungsverfahren nur für einen Teil der Ontologie angewandt wird und die anderen Teile als thematisch weiter entfernt betrachtet werden. Die per Los aus der anzureichernden Ontologie entfernten Kandidaten können hingegen zur Bestimmung der innerontologischen Anreicherungsneigung herangezogen werden. Wenn die innerontologische und fremdontologische Anreicherungsneigung bestimmbar sind, dann deuten wir eine fremdontologische Anreicherungsneigung, die im Vergleich zur innerontologischen zu hoch ausfällt, als negativ, weil zu viele unbrauchbare Begriffe vorgeschlagen werden. **Der Vergleich der innerontologi-**

schen und der fremdontologischen Anreicherungsneigungen dient uns als Hinweis auf die Präzision (Precision, siehe [71]) von Ontologieanreicherungsverfahren. Liegt der Wert der fremdontologischen Anreicherungsneigung absolut niedrig und auch relativ zur innerontologischen Anreicherungsneigung niedrig, dann nehmen wir eine hohe Präzision an. Eine Automatisierung der Vorabbeurteilung von Kandidaten, die bei dieser Betrachtung notwendig ist, werden wir bei den Messungen im nächsten Kapitel liefern.
Abschließend erwähnen wir die Möglichkeit, alle in diesem Abschnitt vorgestellten Maße auch im Mittel über Ontologien zu betrachten. Für die Herstellung aussagekräftiger Stichproben ist die Betrachtung pro Begriff oftmals hilfreich. Wir werden bei den Messungen des folgenden Kapitels auf beide Betrachtungsweisen zurückgreifen.

5.5 Bemerkung: Eigene Beiträge

Das im vorliegenden Kapitel vorgestellte Ontologieanreicherungsverfahren ist in folgenden zentralen Punkten neuartig:

- Funktionsweise des Verfahrens mit genau einem Bezeichner pro Begriff
- Ähnlichkeitsbezug und bestmögliche Vereinbarkeit ontologischer und vektorwertiger Vergleichsmaße durch ein Minimierungsproblem
- Eingangskriterien für Begriffsvorschläge durch die auf die Ontologiestruktur bezogene Definition von Schranken
- automatische Bewertbarkeit der Anreicherungsgüte auch mit genau einem gegebenen Bezeichner pro Begriff

Im technischen Detail wurden neu definiert:

- eine vektorwertige Variante des Jaccardmaßes
- zwei vektorwertige Varianten des Leeschen Schiefmaßes
- Formulierungen der Unähnlichkeit, basierend auf Resniks und Lis ontologischen Vergleichsmaßen
- das asymmetrische ontologische Vergleichsmaß
- Anreicherungsneigungen, Recall- und Fehlklassifikationsmaße für Ontologieanreicherungsprobleme

Diese neuartigen Formulierungen werden im folgenden Kapitel systematischen Tests unterzogen.

Kapitel 6

Implementierung und Messergebnisse

Im folgenden Kapitel stellen wir zunächst in groben Zügen die technische Umsetzung einer Testumgebung für die automatische Anreicherung von Ontologien vor. Dabei beschreiben wir neben einer Architektur für den eigentlichen, im vorangegangenen Kapitel entwickelten Algorithmus mehrere technische Voraussetzungen und Randbedingungen der Umsetzung, die in die Testumgebung integriert werden mussten. Die Details der in der Programmiersprache Perl verfassten Skripte zur Durchführung des Anreicherungsalgorithmus sind dem Anhang zu entnehmen.
Der zweite Schwerpunkt des folgenden Kapitels liegt bei der Präsentation von Messergebnissen. Anhand von Ausschnitten aus der k-med-Ontologie bewerten wir für ein bestimmtes Szenario verschiedene Alternativen des ähnlichkeitsbasierten Ontologieanreicherungsverfahrens. Wir zeigen prinzipielle Wirkungsweisen des Algorithmus und geben entlang unserer quantitativen Ergebnisse einen Vorschlag zur Durchführung weiterer konkreter Ontologieanreicherungen während des Ontologieerstellungsprozesses.

6.1 Datenvorverarbeitung

Die eigentliche automatische Formulierung der Anreicherungsalgorithmen 2 und 3 wurde in eine Umgebung eingebettet, die möglichst vollständig auf freier oder Open Source Software basieren sollte. Wir stellen in den folgenden Abschnitten die Elemente dieser vorhandenen Infrastruktur zur Datenvorverarbeitung und zur mathematischen Umsetzung des Anreicherungsalgorithmus vor.

Die Formate und Wirkungsweisen der Datenvorverarbeitung bestimmen weite Teile der eigentlichen Testimplementierung, so dass wir die folgenden Erläuterungen vor der Architektur der Testumgebung liefern müssen. Insbesondere betrifft die Datenvorverarbeitung die

- Bestimmung von Kollokationen für die Begriffe und Kandidaten
- Eingabe der Struktur der Ontologie, die einer Anreicherung unterzogen werden soll
- die Formulierung des Minimierungsproblems 5.4

Diese einzelnen Charakteristika werden in den folgenden Unterabschnitten dargestellt.

6.1.1 Kollokationsbestimmung

Der folgende Unterabschnitt erläutert die Grundzüge der Kollokationsbestimmung, wie sie in der Testumgebung vorgenommen wurde.
Concollate ist ein Text-Mining-Werkzeug, das für die Datenvorverarbeitung und Textkorpusanalyse im Rahmen der in dieser Arbeit vorgestellten Forschungsarbeiten entwickelt wurde [25]. Concollate dient dazu, in einem Textkorpus für beliebige Begriffe oder Kandidaten Kollokatoren im Sinne der Definition 7 zu finden.
Für die Erstellung des zu analysierenden Textkorpus bestehen zwei Möglichkeiten. Entweder untersucht Concollate einen auf einem lokalen Datenträger festgehaltenen Textkorpus nach Wörtern und ihren Kollokatoren, oder Concollate erzeugt automatisch einen Textkorpus, indem es die zu untersuchenden Wörter (in unserem Sinne: Begriffsbezeichner) als Anfrage an eine WWW-Suchmaschine übergeben lässt und die gefundenen Dokumente speichert.
Das Skript Concollate liefert Daten über Kollokatoren vorgegebener Wörter. Wir erhalten also für eine zu untersuchende Ontologie für jeden Begriff eine Datei mit Kollokationsdaten zum Bezeichner des Begriffes.
In Abbildung 6.1.1 ist die von Conollate erzeugte Datei für einen vorgegebenen Begriff (*Atemwegsinfektionen*) und für einen Abstand von maximal fünf der Kollokatoren (maximales $\delta_W = 5$ im Sinne des Beispiels aus Abschnitt 5.3.1) festgehalten. Hier steht am Anfang jeder Zeile der im Textkorpus gefundene Kollokator. Darauf folgen die Auftrittshäufigkeiten für jede Position vor dem Begriffsbezeichner. An zentraler Stelle steht der Begriffsbezeichner selbst, gefolgt von den Auftrittshäufigkeiten für jede Position hinter dem

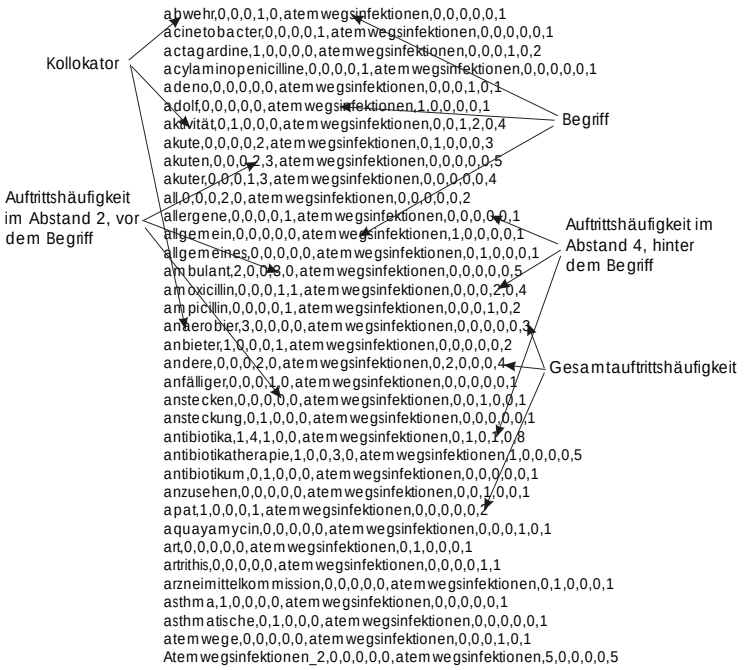

Abbildung 6.1: Aufbau einer Kollokationsdatei

Begriffsbezeichner im Korpus, sowie am Ende jeder Zeile die Gesamtauftrittshäufigkeit des Kollokators mit dem Begriff im Abstand von maximal fünf Wörtern (wiederum maximales $\delta_W = 5$ im Sinne des Beispiels aus Abschnitt 5.3.1).
Die von Concollate erzeugten Dateien werden nach dem untersuchten Begriff benannt und mit der Endung *.con versehen. Die Dateien für alle Begriffe der zu untersuchenden Ontologie müssen in einem separaten Verzeichnis, dem so genannten Con-Verzeichnis auf dem Zielsystem gespeichert werden. Mit diesem Verzeichnis arbeitet dann die Umsetzung des Anreicherungsalgorithmus.
Die Testumgebung ist auf variable Abstände δ_W der Kollokatoren ausgerichtet, die oben genannte von Concollate erzeugte Struktur muss jedoch eingehalten werden. Concollate nimmt keine semantische oder syntaktische Vorverarbeitung der gefundenen Kollokatoren vor, sondern wertet lediglich gemeinsames Vorkommen von Zeichenketten in Texten aus. Das sprachliche Vorwissen, welches hier Verwendung findet, ist somit gering und dem Benutzer wird im Gegensatz zu automatischen Ontologieerstellungsumgebungen wie TextStorm und Clouds (siehe Kapitel 4.1.5) an dieser Stelle keine zusätzliche Pflege oder Wissensrepräsentation in Form von Grammatiken oder Parserregeln abverlangt.

6.1.2 Struktur der Ontologie

Der folgende Unterabschnitt zeigt, wie die Struktur der Ontologie, die einer Anreicherung unterzogen werden soll und die Definition 4 erfüllt, als Eingangsdatei für den Anreicherungsalgorithmus dargestellt werden kann.
Hierfür muss die zu verwendende formale Beschreibung der Ontologie definiert werden. Wir greifen an dieser Stelle wieder die abstrakten Anforderungen aus den ersten Kapiteln der vorliegenden Arbeit auf: die Beschreibung der Ontologie sollte idealerweise auch für Nichtexperten der Wissensrepräsentation den Umgang mit Aufbau und Veränderungen der Ontologie ohne großen Schulungsaufwand möglich werden lassen. Es sollte ohne großen Arbeitsaufwand möglich sein, neue Begriffe an die Ontologie anzufügen, oder für Folgeexperimente unerwünschte Begriffe zu entfernen. Die Ontologiebeschreibung muss darüber hinaus für eine Ontologie eindeutig interpretierbar und maschinenlesbar sein.
Die in der Implementierung verwendete Beschreibung verwendet einen einfachen Zahlenkode, um die Position eines Begriffs in der Ober- und Unterbegriffsstruktur der Ontologie eindeutig zu beschreiben. Die Begriffe werden vom abstrakten Wurzelbegriff $\top$ ausgehend entlang der Geschwister (siehe

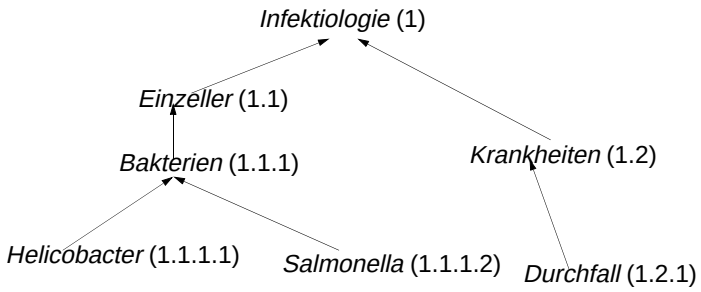

Abbildung 6.2: Nummerierung der Begriffe

Definition 13) für jeden gemeinsamen Oberbegriff durchnummeriert, wobei in jeder neuen Abstraktionsebene eine neue, durch einen Punkt getrennte Stelle eingeführt wird. In Abbildung 6.2 wird der Sachverhalt veranschaulicht, indem die Ontologiestruktur mit mehrstelligen Nummern versehen wird, wobei jede Nummer eine Abstraktionsebene im Sinne der Ausführungen zu den ontologischen Vergleichsmaßen kodiert.

Die Eingangsdatei muss logisch (mittels der Nummerierungen) in Baumform vorliegen. Eine visuell direkt erfassbare Baumstruktur durch Einrückungen wie in Abbildung 6.2 ist nicht erforderlich.
Die Eingangsdatei der Implementierung für dieses Beispiel stellt sich dar wie in Abbildung 6.3.

Definitionsgemäß muss innerhalb der Datei eine Reihenfolge eingehalten werden: Der Zahlenkode eines Begriffes, der ab hier Positionszahl genannt wird, steht an erster Stelle, gefolgt von dem zugehörigen Begriffsbezeichner. Als Trennsymbol wird ein Semikolon verwendet. Die Reihenfolge der Zeilen hingegen ist beliebig, die Begriffe müssen also nicht der Hierarchie des Baumes folgend in die Datei eingetragen werden. Ein neu hinzugefügter Begriff kann daher einfach ans Ende der Datei angefügt werden, unabhängig von

```
1;T
1.1;Einzeller
1.2;Krankheiten
1.1.1;Bakterien
1.2.1;Durchfall
1.1.1.1;Helicobacter
1.1.1.2;Salmonella
```

Abbildung 6.3: Nummerierte Begriffe als Dateiformat

seiner Position in der Ontologie. Eine Ausnahme ist hier ein Begriff, der oberhalb des bisherigen Wurzelknotens stände. Wenn ein solcher Begriff neu in die Ontologie eingefügt wird, müssen die Positionszahlen aller anderen Begriffe von vorn um eine Stelle erweitert werden. Die Bezeichner der Begriffe müssen voll ausgeschrieben und korrekt in die Datei eingetragen sein, da das Programm die Begriffe alphabetisch ordnet und somit ein Rechtschreibfehler beim Erstellen der Datei Auswirkungen auf das Ergebnis hat. Eine weitere syntaktische Besonderheit, die das Dateiformat stellt, liegt darin, dass die Beschreibung der Struktur erst in der fünften Zeile der Datei beginnt. Die ersten vier Zeilen können leer bleiben oder mit Kommentaren bezüglich der beschriebenen Ontologie versehen werden.

6.1.3 Einbindung von AMPL

Der ähnlichkeitsbasierte Anreicherungsalgorithmus stellt auf die Formulierung neuer ontologiespezifischer Vergleichsmaße ab. Dies wird über das Minimierungsproblem 5.4 erreicht und muss ebenfalls in die Testumgebung eingebunden werden.
Die Lösung des Minimierungsproblems 5.4 wird von der Optimierungssprache- und Umgebung AMPL (Akronym für: A Mathematical Programming Language) vorgenommen [29]. Dies erfordert die in Abschnitt 6.1.1 eingeführten Kollokationsdateien und die in Abschnitt 6.1.2 erklärte Ontologiestruktur. Beide zusammen werden zu einem AMPL-spezifischen Format weiterverar-

beitet. Die Implementierung muss also nicht nur die Daten berechnen und intern verarbeiten, sondern sie auch formatieren.
Eine Datei, wie sie von AMPL als Eingangsdatei gefordert wird, ist in Abbildung 6.4 dargestellt. Der Kopf der Datei enthält Resultate des ontologischen Vergleichsmaßes, der Rest der Datei Ergebnisse der vektorwertigen Vergleiche. Die Ontologie, die diesem Beispiel zugrunde liegt, enthält fünf Begriffe. Dies spiegelt sich in den Zuweisungen für I und J wider. L entspricht der Größe der Gewichtung $|\mathcal{A}|$ aus Definition 9.
Im oberen Drittel steht die 5×5-Matrix mit den Werten des verwendeten ontologischen Vergleichsmaßes zwischen den Begriffen.
Die Nummerierung der Matrix bezieht sich auf die alphabetische Abfolge der Begriffe und **nicht** auf die begriffliche Ordnung innerhalb der Ontologie.
Unterhalb der Werte des ontologischen Vergleichsmaßes beginnen die berechneten Vektorentfernungen. Jeder Begriffsvektor wird mit sich selbst und jedem anderen Begriffsvektor durch ein vordefiniertes vektorwertiges Vergleichsmaß im Sinne der Definition 10 verrechnet. Daraus ergeben sich in diesem Fall insgesamt $5^2 = 25$ Vektoren, die jeweils L Einträge aufweisen. Diese Einträge entsprechen den Komponenten des Resultates vektorwertiger Vergleiche. Aus Platzgründen wurde nur ein Ausschnitt der Datei abgebildet. Die Matrix setzt sich nach

```
[*,*,2]
```

analog mit Blöcken bis einschließlich

```
[*,*,5]
```

fort.
Neben der besprochenen Datei benötigt AMPL noch eine .mod-Datei, deren Einträge die zu berechnenden Variablen, also die Zielgröße und den Lösungsvektor des Minierungsproblems 5.4, festlegen. Eine .mod-Datei besitzt die folgende Gestalt

```
 set I; # Index der
Zeilen set J; # Index der Spalten set L; # Index der W"orter

param p >= 0; # Totalzahl der W"orter param dist {I,J} >= 0; #
Rechnung der Distanzen param minima {I,J,L} ; # Rechnung der
Minima
```

```
set I := 1 2 3 4 5 ;
set J := 1 2 3 4 5 ;
set L := 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 ;
param p := 17;
param dist : 1  2  3  4  5   :=
1 0.444444 1.777778 1.777778 1.777778 1.777778
2 2.777778 1.777778 7.111111 7.111111 7.111111
3 2.777778 7.111111 1.777778 7.111111 7.111111
4 2.777778 7.111111 7.111111 1.777778 7.111111
5 2.777778 7.111111 7.111111 7.111111 1.777778
;
```

ontologisches Vergleichsmaß

```
param minima :=
[1,*,*]: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17:=
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 7.46077105556078e-009 -0.0129272100943646 -0.008618150796433
0.0086181507964338
3 0.443094376585252 -0.0465284629701597 -0.0036041967390740
0.00360419673907405
4 9.78933789932238e-009 -0.014082192185761 -0.0093881396970079
0.00938813969700798
5 0.178094024863345 -0.00298766007413901 -0.0319911539514238
[2,*,*]: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17: 17:=
                              ...
                              ...
```

vektorwertiges Vergleichsmaß

Abbildung 6.4: Die AMPL Dateiformate für das Optimierungsproblem

```
var K {L} ; # Werte zu optimieren

minimize werte : sum {i in I} sum {j in J} (dist[i,j]-(sum {l in
L} (K[l]*minima[i,j,l])))*(dist[i,j]-(sum {l in L}
(K[l]*minima[i,j,l])));
```

Hier werden die verschiedenen Indices und Parameter definiert, sowie die zu berechnende Variable K. Die Variablen I, J und L entsprechen den gleichen Variablen in der AMPL-Matrixdatei, die Rechenvorschrift, die durch

```
minimize werte:
```

eingeleitet wird, stellt das Minimierungsproblem aus Gleichung 5.4 dar.

6.2 Testumgebung

Der folgende Abschnitt liefert einen Überblick über die Struktur der Testumgebung.
Ursprünglich erfolgten Design und Programmierung der Testumgebung ohne graphische Benutzerschnittstelle. Daher erscheint es ratsam, die verschiedenen Teile der Umsetzung in unterschiedlichen Skripten, also unterschiedlichen Dateien zu implementieren, um einen effektiveren und spezifischeren Aufruf auch über die Kommandozeile zu ermöglichen. In Abbildung 6.5 sind die Beziehungen der verschiedenen Skripte in einem Komponentendiagramm aufgeführt [1].

Die Testumgebung besteht aus den Skripten Vektorberechnung.pl, Conform.pl, Ontologieberechnung.pl, Kandidatenberechnung.pl und Subs.pl als Teil von Ontologieberechnung.pl. Die Gesamtanwendung wird durch Start.pl gestartet und gesteuert.
Die Interaktion zwischen den einzelnen Skripten und Start.pl erfolgt über Systemaufrufe, der Benutzer interagiert über die grafische Benutzerschnittstelle. Erläuterungen zu den einzelnen Skripten und ihre Funktionalität finden sich im Anhang. Die Abfolge einer typischen Sitzung mit der Testumgebung kann der Abbildung 6.6 entnommen werden. Nach dem Start der Anwendung über das Skript Start.pl werden zunächst vom Benutzer die Eingaben für die zu berechnenden Ontologie, also die Datei mit der Ontologiestruktur und das Verzeichnis mit den Kollokationsdateien, getroffen. In diesem Beispiel wird nun der optionale Vorgang zum Bereinigen

[1] Als Einführung in UML-Komponentendiagramme und -Sequenzdiagramme sei auf [75] verwiesen.

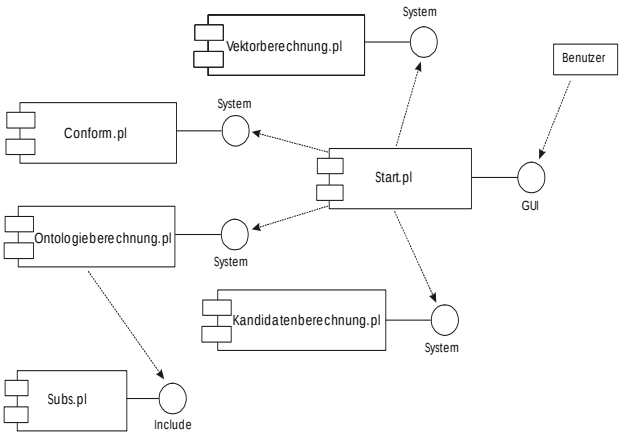

Abbildung 6.5: Komponenten der Testumgebung

des eingegebenen Con-Verzeichnisses gestartet. Das Skript Conform.pl, das die unerwünschten Kollokatoren einer Stopliste aus den Dateien im Con-Verzeichnis entfernt, wird hierzu in einem externen Prozess ausgeführt. Die Stopliste enthält typischerweise Wörter, die als Wort des gegebenen Korpus nach Resnik einen sehr niedrigen Informationsgehalt besitzen [84]. In unserem Falle umfasst dies deutschsprachige Präpositionen, Konjunktionen, Pronomen, Artikel und Hilfsverben.
Im Sequenzdiagramm haben wir für die Bereinigung der Con-Dateien den Terminus 'säubern' und für das bereinigte Verzeichnis den Terminus 'sauber' gewählt.
Sobald die Bereinigung von unerwünschten Kollokatoren beendet ist, kann die Berechnung vom Benutzer initialisiert werden. Hierfür werden hintereinander die dazu nötigen Skripte in ausgelagerten Prozessen aufgerufen. Wenn der erste Teil der Berechnung beendet ist, wird das Zwischenergebnis (Datei provout.dat) als AMPL-Datei, die dem oben erläuterten Format entspricht, angezeigt. Die Bestimmung einer optimalen Gewichtung nach Abschnitt 5.3.2.2 unter Berücksichtigung der Minimierungsvorschrift 5.4 erfolgt anhand von provout.dat.

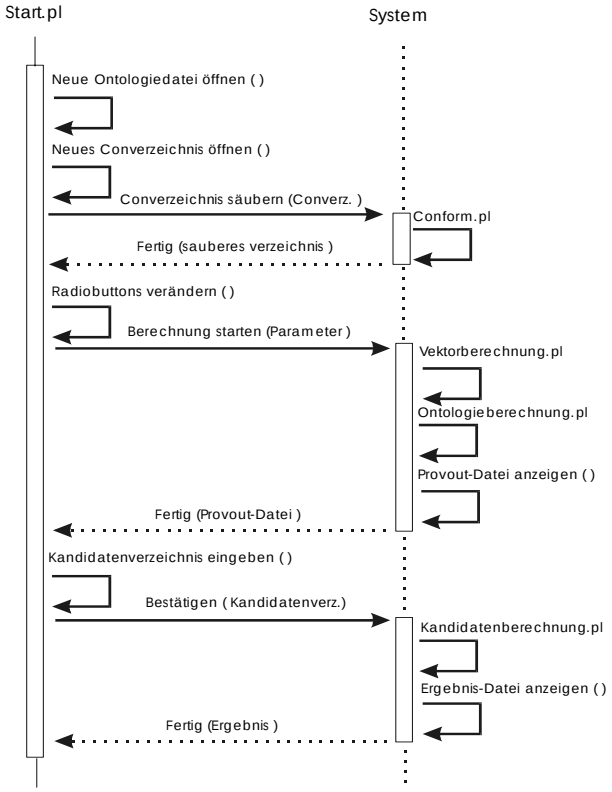

Abbildung 6.6: Anwendungssitzung

Im Folgenden wird der Benutzer dazu aufgefordert, ein Verzeichnis mit Kollokationsdateien möglicher Kandidaten anzugeben. Sobald das Verzeichnis bestätigt ist, werden die für die Beurteilung der Kandidaten nötigen Vorgänge initialisiert. Nach der Anzeige des Endergebnisses, bestehend aus reellen Ähnlichkeits- oder Unähnlichkeitswerten zwischen jedem Begriff und jedem Kandidat, ist der Durchlauf beendet. Optional kann das Ergebnis noch in einer Datei gespeichert werden.
Details zu den einzelnen Perl-Skripten werden im Anhang erläutert.

6.3 Messungen

Im verbleibenden Teil des Kapitels wird die vorgestellte Implementierung zu Messungen mit bestehenden Ontologien herangezogen. Ziel der Messungen ist die Erarbeitung eines normativen Vorschlages zur Verfahrensweise der ähnlichkeitsbasierten Ontologieanreicherung. Dieser Vorschlag soll eine Aussage zu konkreten Anreicherungsszenarien liefern.
Die Unterabschnitte erläutern die Erstellung der für die Messungen verwendeten Daten und die Ergebnisse der Messungen. Schließlich wird durch die Diskussion der Messergebnisse der normative Vorschlag zur Umsetzung des Ontologieanreicherungsverfahrens herausgearbeitet.

6.3.1 Datenbestand

Als Grundlage der Versuche wurde die in k-med [52] entstandene Ontologie, deren Entstehung schon im Kapitel zu den Ontologieerstellungsprozessen erläutert wurde, herangezogen.
Insgesamt wurden sechs verschiedene Fachgebiete der Ontologie ausgewählt. Aus den Bereichen Dermatologie, Infektiologie, Molekularbiologie, Biochemie von Organen und Gewebe, Anatomie von Kopf und Hals, sowie Anatomie der Nerven wurden jeweils zehn bis zwanzig Begriffe ausgewählt. Diese Auswahl erfolgte ohne jegliche Vorinformation über die spätere Gestalt des Textkorpus und ohne jegliche Vorinformation über spätere Anreicherungsmöglichkeiten.
Die Begriffe aus den sechs verschiedenen Fachgebieten wurden so ausgewählt, dass die jeweiligen Graphen, welche die Begriffe als Ecken und die direkten Unterbegriffsrelationen als Kanten haben, zusammenhängend im Sinne von [99] ausfallen:

Definition 14 (Zusammenhang) *Ein Graph $G = (V, E)$ wird zusammenhängend genannt, wenn es für alle Paare von Ecken $v_n, v_m \in V$ einen*

Pfad von v_n nach v_m entlang von Kanten aus E gibt.

Die Beschränkung auf jeweils zehn bis zwanzig Begriffe wurde wegen möglicher, im Vorfeld nicht absehbarer Begrenzungen der Darstellungsmatrizen der eingesetzten Optimierungsanwendung AMPL gewählt. Die Begrenzung betrifft hier die Mächtigkeit der Attributmenge $\mathcal{A}$; bei $|\mathcal{A}| > 300$ wird von der eingesetzten AMPL-Version keine Lösung mehr berechnet.
Für dem grundlegenden Begriffsbestand wurden mit Hilfe von Concollate aus den jeweils ersten 25 Suchtreffern der Suchmaschinen Google und Medivista Kollokationsinformationen generiert. Mit Hilfe von Concollate wurden die Kollokationen im Abstand $\delta_W = 5$ zu jedem Begriffsbezeichner in den für ihn von der Suchmaschine zurückgelieferten Dokumenten ermittelt. Die HTML-spezifischen Teile der Webdokumente wurden entfernt. Das Verfahren zur Konstruktion der Repräsentationsmatrix lief dabei ab, wie im Abschnitt 5.3.1 erläutert. Als Attributmenge $\mathcal{A}$ dienten analog zu unserer dort getroffenen Konstruktion Attribute 'Y kam im Dokument im maximalen Abstand von 5 zum Begriffsbezeichner X vor'. Die Wörter Y, die als Mitglieder von $\mathcal{A}$ überhaupt Berücksichtigung fanden, mussten mit mindestens zwei Begriffsbezeichnern aus der Ontologie im Abstand $\delta_W = 5$ Kollokationen bilden, um Verzerrungen durch ungewöhnliche Kollokatoren zu verringern. Die Wahl fiel auf den Abstand $\delta_W = 5$, da sich nach den Untersuchungen von [85] bis zum Abstand von 5 die meisten Wortassoziationen, die Menschen zu einem vordefinierten Wort angeben, in Textkorpora wiederfinden. Zur Glättung und Definierbarkeit der Vektoreinträge wird analog zur Vorgehensweise in [60] zu jeder Koordinate der Begriffsvektoren ein kleiner Wert von 10^{-7} hinzu addiert.
Für die eigentlichen Anreicherungsexperimente wurden per Los zum einen Ontologien Ω_i' mit jeweils maximal fünf Begriffen ($3 \leq |B| \leq 5$), zum anderen Ontologien Ω_i^* mit fünf bis zehn Begriffen ($5 \leq |B| \leq 10$) hergestellt. Die Ontologien wurden in ihrer kleinen und großen Variante dem Anreicherungsverfahren in den Kombinationen, die im Abschnitt 5.4.2.4 abgeleitet wurden, unterzogen. Die Begriffe, die durch das Los entfernt wurden, dienen nach dem Evaluationsverfahren aus 5.4.3 jeweils als Kandidatenmenge C_i.
Die reellwertigen Schranken T_b, die bei der ähnlichkeitsbasierten Ontologieanreicherung überschritten (im Falle von Unähnlichkeitsmaßen unterschritten) werden müssen, wurden wie in 5.8 als Durchschnittswerte berechnet.
Bezüglich der Wiederanreicherung der Kandidaten C_i wird auch in den Experimenten der 1-Kanten- und der 2-Kanten-Recall (siehe Gleichung 5.35) berechnet. Dies ist anhand der Dermatologie-Ontologie in den Abbildungen 6.7 und 6.8 dargestellt.

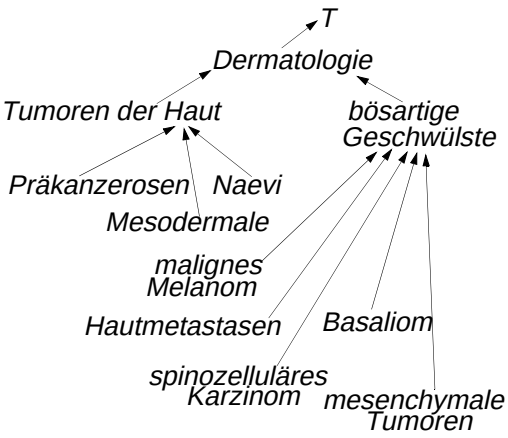

Abbildung 6.7: Beispiel einer ursprünglichen Ontologie

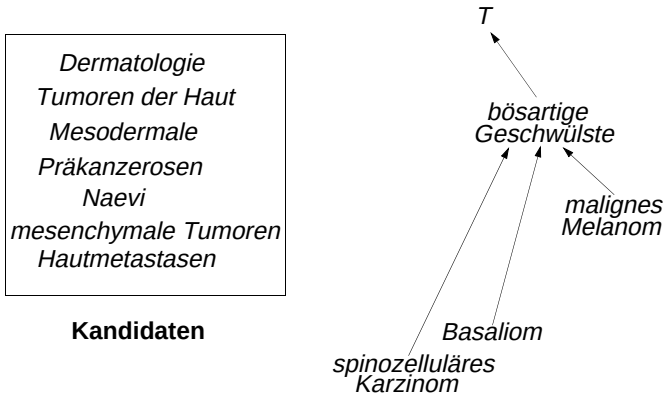

Abbildung 6.8: Ausgeloste Kandidaten und anzureichernde Ontologie

Aus der gegebenen Ontologie in 6.7 wurde per Zufall die Kandidatenmenge, welche in Abbildung 6.8 links zu sichtbar ist, entfernt. Vorschläge aus dieser Menge und ihre Positionierung, die im günstigen Fall wieder ähnlich zur ursprünglichen Dermatologie-Ontologie in 6.7 ausfällt, dienen als Grundlage der Recallmaße.

Abbildung 6.9 zeigt für das Beispiel der anzureichernden Dermatologie-Ontologie Ω' (Unterbegriffe von *Dermatologie*) ein Szenario zur Bestimmung der Anreicherungsneigungen. Eine n-Kanten-Fehlklassifikation bei der Anreicherung von Ω' und die Anreicherungsneigung mit Kandidaten aus Ω, wie wir sie oben per Los hergestellt haben, kann mit der fremdontologischen Neigung zur Anreicherung von Kandidaten aus C (Unterbegriffe von *Infektiologie*, *Anatomie* und *Molekularbiologie*) verglichen werden. Dies wird Bestandteil der folgenden Messungen sein, wobei wir aus verschiedenen anderen Teilen der Gesamtontologie zufällig Kandidaten c mit $c \not\leq \mathit{Dermatologie}$ auswählen. Ein paarweiser Vergleich betrifft dann immer die innerontologische Anreicherungsneigung pro Begriff und die fremdontologische Anreicherungsneigung mit dem selben Begriff.

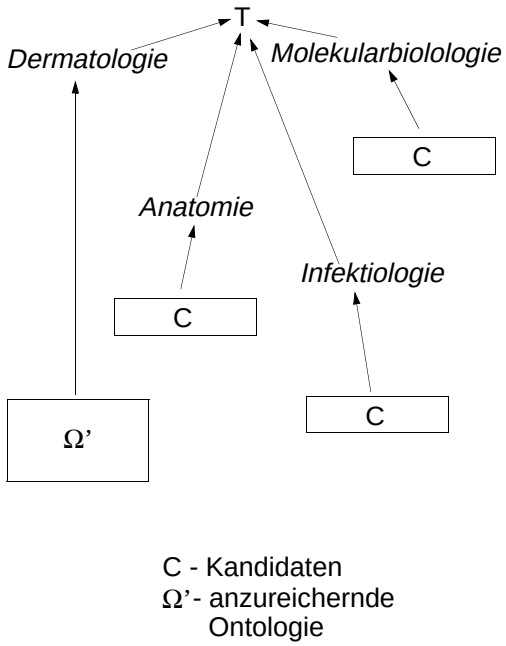

Abbildung 6.9: Auswahl von Kandidaten zur Bestimmung der fremdontologischen Anreicherungsneigung

6.3.2 Naive Anreicherungsstrategie

Um nicht nur einen absoluten Vergleich bezüglich einer vorgegebenen quantitativen Anreicherungsgüte einerseits und einen relativen Vergleich der einzelnen Kombinationen vektorwertiger und ontologischer Vergleichsmaße erreichen zu können, wurde noch das naive Anreicherungsverfahren aus Algorithmus 1 in Anlehnung an die Arbeiten von Heyer [42] auf die Ontologien angewandt.
Im naiven Anreicherungsverfahren, das hier als Referenzverfahren dienen soll, werden die Kollokationen im maximalen Abstand $\delta_W = 5$ direkt als Begriffsvorschläge betrachtet, wie es im Algorithmus 1 in Kapitel 5 erläutert wurde.
Für die durchgeführten Messungen wurde eine obere Schranke des 1-Kanten-Recalls und des 2-Kanten-Recalls für diese naive Anreicherungsstrategie bestimmt. Dies liegt darin begründet, dass die Kollokationsauswertung mit Hilfe von Concollate nur einzelne Wörter als Kollokatoren zu den Begriffsbezeichnern berücksichtigt. Kamen die einzelnen Bestandteile eines aus mehreren Wörtern bestehenden Bezeichners für einen Kandidaten in der Kollokationsdatei vor, so wurde dies optimistisch als Vorkommnis und somit Begriffsvorschlag mit dem zusammengesetzten Bezeichner gewertet. Ebenso wurden bei der Beugung von Wörtern durch Deklination und Konjugation die gebeugten Formen des eigentlichen Bezeichners ebenfalls als Begriffsvorschläge gewertet. Es reichte somit beispielsweise aus, wenn die Kollokatoren 'cystischer' und 'Fibrose' einzeln an verschiedenen Stellen des Textkorpus vorkamen, um 'cystische Fibrose' als Begriffsvorschlag einzuführen.

6.3.3 Ergebnisse

Die herangezogenen Gütemaße sind vom Typ n-Kanten-Recall, n-Kanten-Fehlklassifikation und Anreicherungsneigung, wie sie in Abschnitt 5.4.3 definiert wurden.
Die einzelnen Messungen wurden einem paarweisen Wilcoxon-Test unterzogen, um die Signifikanz der Ergebnisse unabhängig von einer bestimmten Verteilung untersuchen zu können [72][1]. Es lassen sich folgende Ergebnisse statistisch fundiert festhalten.
Der erste Zug der Experimente betraf die kleineren Ontologien. Um sicherzustellen, dass die Ergebnisse nicht zu stark durch die spezifische Wirkungsweise einer Suchmaschine verzerrt werden, wurden aus den insgesamt zwölf

[1] Wilcoxon-Tests untersuchen die Ränge der Werte einer Stichprobe. Der paarweise Wilcoxon-Test untersucht die Ränge der Differenzen der paarweise vorliegenden Werte.

Tabelle 6.1: Legende für die Maßkürzel

Kürzel	*Maß*
S_v	vektorwertiges Vergleichsmaß
S_Ω	ontologisches Vergleichsmaß
J	modifiziertes Jaccardmaß
D	Schiefmaß
D^g	geglättetes Schiefmaß
S^0_Ω	Ähnlichkeitsmaß nach Resnik
S^1_Ω	Ähnlichkeitsmaß nach Li
D^0_Ω	Unähnlichkeitsformulierung des Resnik-Maßes
D^1_Ω	Unähnlichkeitsformulierung des Li-Maßes
D^2_Ω	Unähnlichkeitsformulierung des asymmetrischen Maßes

Ontologien sechs willkürlich ausgewählt.

Tabelle 6.1 liefert nochmals eine Legende für die Kürzel der untersuchten Vergleichsmaße. Die Indices 0.99, die für die Schiefmaße bei den Experimenten zusätzlich auftreten, kennzeichnen den Parameter α aus 5.4.2.2. Als Basisreferenz geben wir in Tabelle die durchschnittlichen 1- und 2-Kanten-Recallwerte für das naive Anreicherungsverfahren bei den kleineren Ontologien an. Für den vorliegenden experimentellen Aufbau fällt der 1-Kanten-Recall des naiven Anreicherungsverfahrens sehr niedrig aus. Der 2-Kanten-Recall liegt etwas höher.
Tabelle 6.3 zeigt den durchschnittlichen 1-Kanten-Recall für die kleinen Ontologien. Die Unterscheidung zwischen Schiefmaßen zwischen b und c als (b, c) und c und b als (c, b) zeigt an, dass im ersten Fall die Unähnlichkeit

Tabelle 6.2: gemittelte Recallwerte für das naive Anreicherungsverfahren bei kleinen Ontologien

Gütemaß	Ω'_1	Ω'_2	Ω'_3	Ω'_4	Ω'_5	Ω'_6
1-Kanten-Recall	0	0	0	0.25	0	0
2-Kanten-Recall	0	0.5	0.1	0.25	0.182	0

zwischen einem Begriff b aus der Ontologie und einem Kandidat b, im zweiten Fall zwischen Kandidat und Begriff herangezogen wurde.

Für die kleineren Ontologien liegt bei einem Signifikanzniveau von mindestens 0.95 der Median des gemittelten 1-Kanten-Recalls unter Beteiligung des Jaccard-Maßes bei über einem Drittel. Dies gilt sowohl für die Kombination mit dem Maß nach Resnik, als auch für die Kombination mit Lis Maß oder dem asymmetrischen ontologischen Vergleichsmaß. Diese statistisch gesicherte Aussage ist hier trotz des mehrfach auftretenden hundertprozentigen 1-Kanten-Recalls vorsichtig zu formulieren, weil der Wilcoxon-Test empfindlich auf den Ausreißer Ω'_5 reagiert. Die Nichtanreicherung im Falle von Ω'_5 zeigt sich aber durchgehend bei allen Verfahren, die aus der Kombination der ontologischen und vektorwertigen Vergleichsmaße entstehen. Für die Varianten unter der Beteiligung der asymmetrischen vektorwertigen Vergleichsmaße liegen die Ergebnisse unter Beteiligung der Schiefmaße im Median signifikant unterhalb der Ergebnisse für die Varianten mit Jaccard. In mehr als der Hälfte der Fälle findet hier im Sinne des 1-Kanten-Recalls keine Anreicherung der erwünschten Art statt. Dies gilt auch für das eigens definierte asymmetrische ontologische Vergleichsmaß.
Tabelle 6.4 zeigt den gemittelten und gerundeten 2-Kanten-Recall für die kleineren Ontologien. Der gemittelte 2-Kanten-Recall liegt lediglich bei der Kombination Jaccard/Resnik im Median signifikant über einem Drittel. Dies gilt für ein 0.95-Signifikanzniveau nach dem Wilcoxon-Test. Wiederum schneiden die Maße unter der Beteiligung von vektorwertigen Schiefmaßen deutlich schlechter ab. In mindestens der Hälfte der Fälle liefern letztere Verfahren keine Anreicherung im Sinne des 2-Kanten-Recalls. Wiederum finden wir einen Ausreißer bei Ω'_5. Eine mögliche Erklärung ist der deutlich geringere Datenbestand bei der für Ω'_5 durchgeführten Messung. Hier wurden von Medivista nur wenige Suchtreffer erzielt, so dass möglicherweise nicht ausreichende Kollokationsinformationen. Da das naive Anreicherungsverfahren im Gegensatz dazu schon beim einmaligen Vorkommnis eines Kollokators

Tabelle 6.3: gemittelter und gerundeter 1-Kanten-Recall bei kleineren Ontologien (mit $c \in C, b \in B'$)

S_v	S_Ω	Ω'_1	Ω'_2	Ω'_3	Ω'_4	Ω'_5	Ω'_6
J	S^0_Ω	0.333	1	1	1	0	1
J	S^1_Ω	0.667	0.667	1	1	0	1
$D_{0.99}(b,c)$	D^0_Ω	0	0	0	0	0	0
$D_{0.99}(b,c)$	D^1_Ω	0.333	0	0	0.5	0	0
$D_{0.99}(b,c)$	D^2_Ω	0.333	0	0	0.5	0	0
$D_{0.99}(c,b)$	D^0_Ω	0	0	0	0.5	0	0.25
$D_{0.99}(c,b)$	D^1_Ω	0	0	0	1	0	0.25
$D_{0.99}(c,b)$	D^2_Ω	0	0	0	0	0	0
$D^g_{0.99}(c,b)$	D^0_Ω	0.333	0	0	0.5	0	0.25
$D^g_{0.99}(c,b)$	D^1_Ω	0.667	0	0	0.5	0	0
$D^g_{0.99}(c,b)$	D^2_Ω	0.333	0	0	0.5	0	0.25
$D^g_{0.99}(b,c)$	D^0_Ω	0	0	0	0.5	0	0
$D^g_{0.99}(b,c)$	D^1_Ω	0	0.333	0	1	0	0
$D^g_{0.99}(b,c)$	D^2_Ω	0	0	0	0.5	0	0

Tabelle 6.4: gemittelter und gerundeter 2-Kanten-Recall bei kleineren Ontologien

S_v	S_Ω	Ω'_1	Ω'_2	Ω'_3	Ω'_4	Ω'_5	Ω'_6
J	S^0_Ω	0.375	1	1	0.75	0	0.941
J	S^1_Ω	0.5	0.313	1	1	0	1
$D_{0.99}(b,c)$	D^0_Ω	0.125	0.167	0	0	0	0
$D_{0.99}(b,c)$	D^1_Ω	0.188	0	0	0	0	0
$D_{0.99}(b,c)$	D^2_Ω	0.436	0	0	0.25	0	0
$D_{0.99}(c,b)$	D^0_Ω	0	0	0	0.25	0	0.15
$D_{0.99}(c,b)$	D^1_Ω	0	0	0	0.375	0	0.083
$D_{0.99}(c,b)$	D^2_Ω	0.063	0	0	0.125	0	0
$D^g_{0.99}(c,b)$	D^0_Ω	0.188	0	0	0.375	0	0.235
$D^g_{0.99}(c,b)$	D^1_Ω	0.313	0	0	0.125	0	0.118
$D^g_{0.99}(c,b)$	D^2_Ω	0.375	0	0.1	0.125	0	0.057
$D^g_{0.99}(b,c)$	D^0_Ω	0.063	0	0	0.125	0	0.177
$D^g_{0.99}(b,c)$	D^1_Ω	0	0.083	0.1	0.375	0	0
$D^g_{0.99}(b,c)$	D^2_Ω	0	0.25	0	0.625	0	0

Tabelle 6.5: gemittelte und gerundete 2-Kanten-Fehlklassifikation bei kleineren Ontologien

S_v	S_Ω	Ω'_1	Ω'_2	Ω'_3	Ω'_4	Ω'_5	Ω'_6
J	S^0_Ω	0.75	1	nicht definiert	nicht definiert	0	0.93
J	S^1_Ω	0.75	0.375	nicht definiert	nicht definiert	0	1

wirkt, erfolgt bei Ω'_5 eine Anreicherung dadurch, wie Tabelle 6.2 beim 2-Kanten-Recall zeigt.
Für den gemittelten 1-Kanten-Recall ist der Unterschied zwischen den auf dem Jaccardmaß basierenden Anreicherungen und der naiven Anreicherungsstrategie signifikant (Signifikanzniveau mindestens 0.95). Eine solche Signifikanz wird für den gemittelten 2-Kanten-Recall (trotz eines höheren Werts in allen Fällen außer dem Ausreißer Ω'_5) nicht erreicht. Ein Signifikanzniveau von mindestens 0.95 wird jedoch für den Fall, dass man den 2-Kanten-Recall pro einzelnem Begriff vergleicht, erreicht.
Tabelle 6.5 zeigt die gemittelte und gerundete 2-Kanten-Fehlklassifikation für die nach obigen Ergebnissen anreicherungsfreudigen Kombinationen Jaccard/Resnik und Jaccard/Li. Die Messreihe deutet an, dass mit den guten Recallwerten Fehlklassifikationen verbunden sein können. Allerdings liegt hier eine Situation vor, in der der fehlerhaft vorgeschlagene Begriff genau 3 Kanten vom Begriff, zu welchem vorgeschlagen wurde, entfernt war. Wir interpretieren die Messung dahingehend, dass die Fehlklassifikation **innerhalb** eines thematischen Bereiches stattfindet und somit eine Unschärfe vorliegt. Bei den beiden Ontologien, zu denen das Fehlklassifikationsmaß nicht definiert war, lagen bei unserem experimentellen Aufbau keine Kandidaten vor, die mehr als 3 Kanten vom Begriff, zu dem vorgeschlagen wurde, entfernt waren. Diese Situation wiederholt sich bei den großen Ontologien, so dass wir auf einen zusätzlichen Indikator bezüglich des Fehlklassifikationsverhaltens zurückgreifen, wie er der Messung in Tabelle 6.6 zu Grunde liegt. Das Fehlklassifikationsmaß pro Begriff sinkt für die Kombination Jaccard/Resnik signifikant, wenn man pro Begriff den Median der fremdontologischen Anreicherungsneigung mit der innerontologischen Anreicherungsneigung vergleicht. Auch hier gilt ein Signifikanzniveau von mindestens 0.95. Der Median sinkt für die Stichprobe von Begriffen, die zufällig aus den anzureichernden Ontologien ausgewählt wurden, gegenüber denjenigen, die zufällig aus den nicht anzureichernden Ontologien ausgewählt wurden, von 1 auf 0.235, also

Tabelle 6.6: Anreicherungsneigungen pro Begriff für kleine Ontologien, Jaccard/Resnik

Begriff	fremdontologische Anreicherungsneigung	2-Kanten-Fehlklassifikation
b_1	0.091	1
b_2	0.455	1
b_3	0.068	1
b_4	0	1
b_5	0.235	1
b_6	0.373	1
b_7	0.373	1
b_8	0.078	1
b_9	0.392	0.563

b_1 =*Basaliom*, b_2 =*bösartige Geschwülste*, b_3 =*malignes Melanom*, b_4 =*spinozelluläres Karzinom*, b_5 =*Atemwegsinfektionen*, b_6 =*chronische Bronchtis*, b_7 =*Lungenabszess*, b_8 =*Mukoviszidose*, b_9 =*Pneumokokken*

Tabelle 6.7: Durchschnittlicher 2-Kanten-Recall für größere Ontologien, Jaccard/Resnik

S_v	S_Ω	Ω_1^*	Ω_2^*	Ω_3^*	Ω_4^*	Ω_5^*	Ω_6^*
J	S_Ω^0	0.667	0.397	0.444	0.423	0.209	0.167

auf knapp ein Fünftel.

Die größeren Ontologien wurden im zweiten Zug der Experimente untersucht. Hierbei fanden lediglich auf Google basierende Kollokationsinformationen aus sechs Ontologien $\Omega_1^*, ..., \Omega_6^*$ Verwendung. Hauptsächlich verfolgten die Messungen das Ziel, die Eigenschaften der Maßkombination Jaccard/Resnik zu verifizieren. Die durchschnittlichen Werte für den 2-Kanten-Recall dieser Maßkombination finden sich in Tabelle 6.7. Im Folgenden sind die Signifikanzen immer am 0.95-Niveau ausgerichtet.

Die Werte in Tabelle 6.7 liegen im Median signifikant über 0.2 und es findet sich kein Ausreißer, bei dem die Anreicherung nicht mehr stattfindet, darunter. Zudem zeigte eine Stichprobe von 27 Begriffen aus $\Omega_1^*, ..., \Omega_6^*$, dass pro Begriff der 2-Kanten-Recall über 0.32 liegt. Für die Kombination aus Jaccard und Li liegt der gemittelte 2-Kanten-Recall nicht signifikant über 0.2.
Wir zeigen wieder die Ergebnisse für die naive Anreicherungsstrategie durch einen Vergleich zur ähnlichkeitsbasierten Anreicherung mit der Maßkombination Jaccard/Resnik in Tabelle 6.8.

Die naive Anreicherung liegt beim 2-Kanten-Recall pro Begriff mit einem Median zwischen 0 und 0.111 signifikant unterhalb der ähnlichkeitsbasierten Ontologieanreicherung mit dem Median von 0.556 für die Stichprobe von Begriffen aus den sechs größeren Ontologien.
Die 2-Kanten-Fehlklassifikation war aufgrund der gegebenen Ontologiestrukturen zumeist nicht definiert. Tabelle 6.9 zeigt den daher getroffenen Vergleich zwischen innerontologischen und fremdontologischen Anreicherungsneigungen.

Der Median der fremdontologischen Anreicherungsneigung sinkt für eine Stichprobe von Begriffen aus den sechs größeren Ontologien signifikant auf 0. Im Mittel findet somit die Anreicherung fachfremder Kandidaten als Begriffsvorschläge nicht statt.

Tabelle 6.8: 2-Kanten-Recall für größere Ontologien: naive Anreicherung mit Google-Korpus gegen Jaccard/Resnik/Google

Begriff	naiv	Jaccard/Resnik
dura mater spinalis	0	1
arachnoidea mater	0	1
Zentralnervensystem	0	1
pneunomitis carinii pneunomie	0	0
Mukoviszidose	0.111	0.556
cystische Fibrose	0.111	0.556
Ketogenese	0.4	1
Biochemie	0.5	0
Gluconeogenese	0.667	1
Präkanzerosen	0	0
malignes Melanom	0.667	0
mesenchymale Tumoren	0	0.5
Infektiologie	0.111	0.333
Anatomie	0	0
Plasmaprotein	0	0.75

Tabelle 6.9: Anreicherungsneigung pro Begriff in großen Ontologien, Jaccard/Resnik, mit ursprünglichem Ω, anzureicherndem Ω^*, fachfremdem Ω_x

Begriff	innerontologische Anreicherungsneigung, $C \subset \Omega$	fremdontologische Anreicherungsneigung, $C \subset \Omega_x$
b_1	1	0.6
b_2	1	0
b_3	0.667	0
b_4	0.667	0
b_5	0.778	0
b_6	0.556	0.6
b_7	0.333	0
b_8	0.556	0
b_9	0	0
b_{10}	1	0
b_{11}	0.333	0.9
b_{12}	0.833	0
b_{13}	0.833	0
b_{14}	0.833	0
b_{15}	0.833	1
b_{16}	1	0
b_{17}	0	0
b_{18}	0	0.5
b_{19}	0	0
b_{20}	1	0

b_1 =*Arachnoidea mater*, b_2 =*Ausgangsmaterial*, b_3 =*Akute bronchitis*,
b_4 =*Bronchiolitis*, b_5 =*chronische Bronchitis*, b_6 =*cystische Fibrose*,
b_7 =*Infektiologie*, b_8 =*Mukoviszidose*,
b_9 =*pneunomitis carinii Pneunomie*, b_{10} =*Cofaktoren*,
b_{11} =*Gallenflüssigkeit*, b_{12} =*Gluconeogenese*, b_{13} =*Glucosehomöostase*,
b_{14} =*Ketogenese*, b_{15} =*Plasmaprotein*, b_{16} =*Triacylglycerin*,
b_{17} =*Dermatologie*,
b_{18} =*ektodermale*, b_{19} =*malignes Melanom*, b_{20} =*Tumoren der Haut*

6.3.4 Diskussion und Schlussfolgerung

Der folgende Abschnitt gliedert sich in zwei Teile. Die Diskussion der obigen Messergebnisse fasst die wichtigsten Erkenntnisse für das gegebene Szenario zusammen. Die Schlussfolgerung legt dar, inwieweit die experimentelle Situation auf einen Vorschlag zur Anwendung der ähnlichkeitsbasierten Ontologieanreicherung übertragbar ist.

6.3.4.1 Diskussion der Messungen

Die Messergebnisse, die im vorigen Abschnitt gebündelt dargestellt wurden, legen nahe, die vektorwertigen Unähnlichkeitsmaße für die Anreicherung der gegebenen Ontologien zu verwerfen. Die Recallwerte für die kleinen Ontologien erweisen sich auch im Vergleich mit der naiven Anreicherungsstrategie als zu niedrig, um ein positives Anreicherungsergebnis im Sinne einer Vervollständigung der Ontologie zu liefern. Eine vorsichtige Strategie legt nahe, trotz der denkbaren günstigeren Anreicherungsergebnisse für größere Ontologien die aufgeführten Versionen der Schiefmaße nicht weiter zu berücksichtigen. Eine Kombination des Schiefmaßes mit dem asymmetrischen Vergleichsmaß konnte in den Experimenten ebenfalls keine Verbesserung erzeugen.
Prinzipiell stellen im vorgestellten Fall die vektorwertigen Unähnlichkeitsmaße, die auf dem Jaccardmaß beruhen, eine geeignete Möglichkeit zur konkreten Umsetzung des Ontologieanreicherungsverfahrens dar. Insbesondere kann ihre Kombination mit dem Ähnlichkeitsmaß nach Resnik als geeignete Wahl der vektorwertigen und ontologischen Vergleichsmaße angesehen werden. Dies gilt sowohl für die kleineren als auch für die größeren Ontologien, zu denen Messungen durchgeführt wurden.
Für diese Kombination aus dem Jaccardmaß und Resniks Maß kann auch signifikant eine grundsätzlich wünschenswerte Wirkungsweise des ähnlichkeitsbasierten Ontologieanreicherungsverfahren gezeigt werden. Die Anzahl unerwünschter Begriffsvorschläge nimmt nämlich signifikant ab, wenn wir Kandidaten heranziehen, die innerhalb der ursprünglichen Ontologie zu anderen Fachgebieten zu zählen waren (Vergleich inner- und fremdontologischer Anreicherungsneigung). **Die günstigen Recallwerte für die Kombination aus dem Jaccardmaß und Resniks Maß werden somit im Rahmen unserer Messungen nicht durch eine eventuelle unzumutbar starke Neigung des Algorithmus zu vielen Vorschlägen konterkariert.** Dies gilt insbesondere dann, wenn die Kombination des modifizierten Jaccardmaßes und des Resnikmaßes bei den größeren Ontologien mit

auf Google-Suchtreffern basierenden Kollokationsinformationen angewandt wird. Die niedrigeren durchschnittlichen Recallwerte im Vergleich zu den kleineren Ontologien sind in diesem Falle in Kauf zu nehmen, da gleichzeitig wesentlich weniger (im Median 0) fremde Begriffe vorgeschlagen werden. Im Vergleich zu den kleineren Ontologien (Median der fremdontologischen Anreicherungsneigung: 0.235) stellt dies eine sehr wünschenswerte Verbesserung dar.
Andererseits ist das Anreicherungsverfahren dem naiven Anreicherugsverfahren aus Sicht der Recall-Evaluierungen deutlich überlegen. Die mangelnde sprachliche Vorverarbeitung des naiven Anreicherungsverfahrens kann hierbei nicht als Begründung herangezogen werden, da eine entsprechende obere Schranke des Recalls gemessen wurde und die bevorzugte Kombination aus ontologischem und vektorwertigem Vergleichsmaß auch dieser oberen Schranke überlegen war.
Eine mögliche Erklärung im Rahmen für das bessere Abschneiden der ähnlichkeitsbasierten Maße gegenüber den unähnlichkeitsbasierten Maßen kann durch folgende Betrachtung gewonnen werden. Wir ziehen für alle Kandidaten der Anreicherung kleiner Ontologien die mittlere durchschnittliche Abweichung des durch die optimale Gewichtung $\vec{k}_{opt}(\Omega_i')$ gewonnenen Vergleichsmaßes vom tatsächlichen ontologischen Vergleichsmaß heran. Dazu verwenden wir folgende Notation. Sei Ω_i eine der ursprünglich gegebenen Ontologien, bei der noch keine Kandidaten zu Evaluationszwecken entfernt wurden. Es seien Ω_i' eine der kleineren Ontologien mit den Begriffen B_i' und C_i die dazu gehörige Kandidatenmenge. Sei außerdem $\vec{k}_{opt}(\Omega_i')$ die für Ω_i' gefundene optimale Gewichtung. Dann gibt

$$A(\Omega_i, \Omega_i') := \frac{\sum_{b \in B_i'} \sum_{c \in C_i} \left(\frac{|\vec{k}_{opt}(\Omega_i') S_v(c,b) - S_{\Omega_i}(c,b)|}{S_{\Omega_i}(c,b)} \right) + \frac{|\vec{k}_{opt}(\Omega_i') S_v(b,c) - S_{\Omega_i}(b,c)|}{S_{\Omega_i}(b,c)}}{|B_i'||C_i|} \tag{6.1}$$

an, wie stark die mittels der optimalen Gewichtung gefundenen Vergleichswerte im Verhältnis zum ursprünglichen Vergleichswert vom ursprünglichen Vergleichswert abweichen. Wir betrachten den Mittelwert von $A(\Omega_i, \Omega_i')$ über unsere sechs gegebenen Ontologien für den Fall der kleinen Ontologien Ω_i' und erhalten die in Tabelle 6.10 aufgetragenen Werte.

Die Werte für das Jaccardmaß liegen mit 1.099 für Li und 2.351 für Resnik niedriger als die Mittelwerte für die asymmetrischen vektorwertigen Maße. In diesem Szenario erfolgt eine bessere Anpassung an Jaccardmaße nach dem Verständnis der Gleichung 5.4. Für Recallwerte, Anreicherungsneigungen und Fehlklassifikationen gehen zwar jeweils nur ein Teil der in

Tabelle 6.10: Anpassung der vektorwertigen und der ontologischen Vergleichsmaße

	Jaccard	*Schiefmaß*	*geglättetes Schiefmaß*
Resnik	2.351	52.4045	242.15
Li	1.099	7.119	7.166
asymmetrisches Vergleichsmaß	5.283	15.084	5.162

Tabelle 6.10 berücksichtigten Vergleiche ein, trotzdem liefern die Ergebnisse Grund zu der Annahme, dass die über das Minimierungsproblem 5.4 gefundenen Maße für unter der Beteiligung des modifizierten Jaccardmaßes besser für das generelle Vorgehen geeignet sind und besser zur Rekonstruktion der ontologischen Vergleichsmaße durch das Minimierungsproblem 5.4 herangezogen werden können.

6.3.4.2 Schlussfolgerung und Vorschlag

Die Experimente des vorangegangenen Abschnittes zeigen, dass sich in der gegebenen Situation die prinzipielle erwünschte Wirkungsweise der ähnlichkeitsbasierten Ontologieanreicherung mittels einer bestimmten Konfiguration des Algorithmus (Wahl der Maße und Ausgangsgröße der Ontologien) erreichen lässt. Da in anderen Ausgangssituationen andere Ausgangsgrößen (Ontologien und Korpora) vorliegen können, liefern wir im letzten Abschnitt dieses Kapitels einen Vorschlag, wie die systematische Durchführung von Messungen und Anreicherungen im Ontologieerstellungsprozess eingesetzt werden kann.
Wir beziehen uns wieder auf das kooperative Ontologieerstellungsverfahren, wie es in k-med eingesetzt wurde. Die zur Anreicherung notwendigen Aktivitäten des Moderators erfolgen nicht zeitgleich mit der Ontologieerstellung, sondern während die Aktivitäten der Autoren ruhen. Das ähnlichkeitsbasierte Ontologieanreicherungsverfahren kann sowohl in der Verankerungsphase als auch in der Phase der iterativen Verbesserung (die im Falle einer Verschlagwortung einer Wechselwirkung mit der Anwendung unterliegt, siehe Abbildung 3.3) eingesetzt werden. Für beide Phasen gilt, dass die technische Durchführung der Ontologieanreicherung beim Moderator liegen sollte.
Der Moderator führt für eine gegebene Ontologie mehrere Formen der Zerlegung in kleinere zusammenhängende Ontologien durch. Dies geschieht, indem die Gesamtmenge der Begriffe in disjunkte Teilmengen der Begriffe und

der für sie gültigen Ober- und Unterbegriffsrelationen unterteilt wird. Zudem benötigt der Moderator von den Autoren der Ontologie Textkorpora, die als charakteristisch für das zu modellierende Fachgebiet anzusehen sind. Zusätzlich sollten die Autoren auch noch Listen von Kandidaten mitliefern. Dabei kann es sich beispielsweise um das Stichwortverzeichnis eines Lexikons, um den Index eines Fachbuches zum Themengebiet oder um Schlüsselwörter, die der Autor zur Verschlagwortung oder als Metadateneinträge wünscht und die aber noch nicht als Begriffe der Ontologie angelegt sind, handeln. Liegt eine solche Kandidatenliste nicht vor, so kann zwar grundsätzlich jedes Wort aus dem gegebenen Textkorpus daraufhin untersucht werden, ob es zum Begriffsvorschlag des Ontologieanreicherungsverfahrens wird, jedoch wäre eine Identifikation von Bezeichnern, die aus mehreren Wörtern (im Dermatologie-Beispiel des vorliegenden Kapitels: *malignes Melanom*, *spinozelluläres Karzinom* et cetera) bestehen, notwendig.
Mit der Zerlegung der Ontologie wird der Anreicherungsprozeß unabhängig von einer Skalierbarkeit des ähnlichkeitsbasierten Verfahrens im Allgemeinen und der Lösung des Minimierungsproblems 5.4 im Speziellen. Der Moderator prüft für verschiedene Zerlegungen der gegebenen Ontologie ähnlich wie in den vorgestellten Experimenten, wie sich die Recallmaße sowie die inner- und fremdontologischen Anreicherungsneigungen gestalten. Dazu müssen per Los Teile der aus der Zerlegung resultierenden Ontologien als Kandidaten identifiziert werden. Die Messungen des letzten Abschnittes erheben in diesem Zusammenhang keinen Anspruch auf Vollständigkeit, sondern zeigen, welche Situation der ähnlichkeitsbasierten Anreicherung als günstig identifiziert werden kann; die in der Auswertung unseres Experiments favorisierte Kombination aus dem modifizierten Jaccardmaß und dem Resnik-Maß erfüllt folgende Bedingungen:

- relativ hohe Recallwerte oberhalb der Vergleichswerte für das naive Anreicherungsverfahren
- niedrige fremdontologische Anreicherungsneigung im Verhältnis zur innerontoloigschen Anreicherungsneigung

Bei der Bestimmung der fremdontologischen Anreicherungsneigung muss im Vergleich zu unserem experimentellen Aufbau beachtet werden, dass es durchaus positiv sein kann, wenn der eigene fachgebietsspezifische Korpus eines Autoren Vorschläge liefert, die als Begriffe in anderen Teilen der Ontologie bereits vorhanden sind oder als Begriffsvorschläge in anderen Teilen der Ontologie auftauchen. Die fremdontologische Anreicherungsneigung sollte daher mit Korpusinformationen aus den fremden Textkorpora berechnet

werden und ist als relatives Maß anzusehen. Das bedeutet, dass die fremdontologische Anreicherungsneigung beim Vergleich verschiedener Zerlegungen und Maßkombinationen gering ausfallen sollte, ein Median von 0, wie er in unseren Messungen nachgewiesen werden konnte, sollte nicht als absoluter Richtwert dienen. Die Korpusgrößen sollten zudem nicht stark voneinander abweichen. Ziel des Vergleiches ist somit die Herstellung optimaler Gewichtungen, die möglichst korpusspezifisch sind. Will sich der Moderator auf Untersuchungen von inner- und fremdontologischen Anreicherungsneigungen mit einem Textkorpus beschränken, so sollte die damit anzureichernde Ontologie hinreichend groß sein, um auch tatsächlich eine Herstellung von als Begriffsvorschlag negativ aufzufassenden Kandidaten möglich werden zu lassen.

Der Moderator wählt eine Zerlegung und eine Maßkombination, die bei seinen Tests als sinnvoll identifiziert wurde, indem ihr eine signifikant bessere Wirkungsweise als anderen Zerlegungen und Maßkombinationen nachgewiesen werden konnte. Mit dieser gefundenen Ausprägung des Anreicherungsansatzes werden schließlich die Begriffsvorschläge zu den vorhandenen Begriffen hergestellt und den Autoren geliefert. Dabei ist zu beachten, dass das Verfahren die Begriffsvorschläge aufgrund derjenigen Ontologien liefert, welche nach Entfernung der Kandidaten resultieren (in unseren Experimenten Ω_i' und Ω_i^*). Tritt ein gleich lautender Begriffsvorschlag an mehreren Stellen der Gesamtontologie auf oder ist er bereits an einer andere Stelle der Gesamtontologie vorhanden, so sind die für die jeweiligen thematischen Bereiche verantwortlichen Ontologieautoren zur Kooperation aufgerufen. Der Moderator sollte diese Überschneidungen den Autoren mitteilen, um die Kooperation zu erleichtern.

Das gesamte Verfahren kann im Erstellungsprozess der Ontologie wiederholt werden.

Kapitel 7

Zusammenfassung

Die vorliegende Arbeit stellt einen neuartigen Ansatz dar, eine automatische Unterstützung des Ontologieerstellungsprozesses zu liefern. Das Verfahren kann bestehende Ontologien erweitern und so einen Beitrag zur erleichterten Konstruktion und Wiederverwendung bestehender Wissensrepräsentationen liefern.
Es wurde dazu eine ähnlichkeitsbasierte Methode definiert und erprobt. Die Wirkungsweise der Methode gleicht die Ähnlichkeiten zwischen Begriffspaaren aus der Ontologie und die Ähnlichkeiten, die sich aus der Verwendung von Fachbegriffen in einem Textkorpus bestimmen lassen, einander an.
Die in der Arbeit dargelegten Messergebnisse zeigen, dass für ein beispielhaftes Szenario eine Definition des Verfahrens gefunden werden kann, die anderen Verfahren signifikant überlegen ist: die Kombination aus dem neu definierten vektorwertigen Ähnlichkeitsmaß, das auf dem Jaccardmaß beruht, und dem ontologischen Vegleichsmaß nach Resnik.
Die Bewertung des Verfahrens im Beispielszenario beruht auf einem Datenbestand, wie er in einer realistischen Projektsituation (k-med) entsteht.
Die in der Arbeit definierte Vorgehensweise der Messungen ist auf weitere konkrete Anwendungsfälle, in denen Ontologieanreicherungen vorgenommen werden sollen, übertragbar.
Alle in dieser Arbeit neu definierten Kriterien der Bewertung, nämlich Fehlklassifikation, Anreicherungsneigung und Recalleigenschaften, stehen in unmittelbarem Zusammenhang mit verschiedenen Phasen des kooperativen Ontologieerstellungsprozesses nach Holsapple und Joshi, dessen iterative Phasen im Rahmen dieser Arbeit mit der Erstellung von Metadaten und noch allgemeiner mit der Verschlagwortung von Dokumenten integriert wurden.

Das ähnlichkeitsbasierte Verfahren eröffnet in zweifacher Hinsicht Möglichkeiten der automatischen Unterstützung. Sind eine Ontologie und ein Textkorpus gegeben, so kann der Moderator des Ontologieerstellungsprozesses jedem Ersteller der Ontologie **unabhängig** von der Anwendung der Ontologie Begriffsvorschläge und ihren Bezug zur bereits vorhandenen Ontologie zukommen lassen. Dies kann im Erstellungsprozess bereits sehr früh und sehr spezifisch, das heißt mit kleinen fachbezogenen Ontologien, geschehen. Die vorliegende Arbeit geht in diesem Sinne über vorhandene Arbeiten der vollständig automatischen Generierung von Ontologien hinaus.
Darüber hinaus kann die Verschlagwortung mit neuen Begriffen während des Ontologieerstellungsprozesses einen Rückbezug auf die Ontologie erfahren, wie es für eine kontinuierliche anwendungsbezogene Erweiterung fachgebietsspezifischer Ontologien gefordert wurde.
Alle Untersuchungen der vorliegenden Arbeit wurden mit dem Anspruch auf eine möglichst starke Ausrichtung an statistischen Verfahren durchgeführt. Es wurde somit die Wirksamkeit eines Verfahrens ohne zusätzliches, aufwändig zu pflegendes sprachliches oder logisches Hintergrundwissen gezeigt. Daher stützen sich die Auswertungen der Textkorpusinformationen zur Verwendung von Fachbegriffen auf Kollokationen von Begriffsbezeichnern und Wörtern in fachspezifischen Textkorpora. Durch die Einführung von Ähnlichkeitswerten und Begriffsvorschlägen bleibt das Ontologieanreicherungsverfahren für Anwender transparent, der letzte Schritt der konkreten Integration eines Begriffsvorschlags obliegt der Gestaltungsfreiheit der Nutzer.
Mit dem in der vorliegenden Dissertationsschrift entwickelten Verfahren wird die Erstellung von Ontologien systematisiert und vereinfacht. In der erzielten Vereinfachung ist ein Beitrag zum Einsatz von Ontologien in Anwendungen der Wissensrepräsentation zu sehen.

Kapitel 8

Literaturangaben

Folgende Veröffentlichungen des Verfassers als Erstautor flossen in die vorgestellte Arbeit ein.

(i) Andreas Faatz und Ralf Steinmetz: Measuring Ontology Enrichment Quality. Kurzpapier (mit Poster), Tagungsband EKAW 2004, Whittlebury Hall, Northamptonshire, Springer Lecture Notes on Computer Science, Oktober 2004

(ii) Andreas Faatz: Semantic Distances and Similarities in Ontology Enrichment , erscheint als Buchkapitel in Acquisition and representation of word meaning, Theoretical and computational perspectives, edited by Alessandro Lenci Universit di Pisa -Dipartimento di Linguistica T. Bolelli (Pisa, Italy), Simonetta Montemagni Istituto di Linguistica Computazionale - CNR (Pisa, Italy), Vito Pirelli Istituto di Linguistica Computazionale - CNR (Pisa, Italy), 2004

(iii) Andreas Faatz und Ralf Steinmetz: Precision and Recall for Ontology Enrichment, in ECAI 2004-Workshop-Tagungsband Ontology Learning and Population, Valencia, August 2004

(iv) Andreas Faatz, Cornelia Seeberg und Ralf Steinmetz: Ein Begriffsnetz für ein medizinisches Online-Lernprojekt. In Proceedings des LEARNTEC 03 WORKSHOPS Aufbau und Einsatz von Begriffsnetzen zur semantischen Suche von Wissensmaterialien, Karlsruhe, März 2003.

(v) Andreas Faatz und Ralf Steinmetz: Statistical Profiles of Words for Ontology Enrichment. In Premyslaw Grzegorzewski, Olgierd Hryniewicz, Maria A. Gil (Herausgeber) In Proceedings of SMPS, Warschau 2002. Springer Serie 'Advances in Soft Computing', 2002.

(vi) Andreas Faatz und Ralf Steinmetz: Ontology Enrichment with Texts from the WWW . In Proceedings of ECML-Semantic Web Mining Workshop 2002, Helsinki, August 2002.

(vii) Andreas Faatz, Stefan Hoermann, Cornelia Seeberg, und Ralf Steinmetz: Conceptual Enrichment of Ontologies by means of a generic and configurable approach., im ESSLLI-Workshop-Tagungsband 'Workshop on Semantic Knowledge Acquisition and Categorisation', Helsinki, August 2001.

(viii) Andreas Faatz, Adulmotaleb El Saddik, Stefan Hoermann, Ivica Rimac, Cornelia Seeberg, Achim Steinacker, und Ralf Steinmetz: Multimedia und Wissen: Unser Weg zu einem produktiven Umgang mit Wissensdurst. In thema Forschung, 2000(2), November 2000.

(ix) Andreas Faatz, Thomas Kamps, und Ralf Steinmetz: Background Knowledge, Indexing and Matching- Interdependencies of Document Management and Ontology Maintenance, Kurzpapier, im Tagungsband des ersten Workshops 'Ontology Learning', ECAI 2000, Berlin, August 2000.

Literaturverzeichnis

[1] E. Aguirre, M. Lersundi: Extracción de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición. In Procesamiento del Language Natural **27**, 2001

[2] D. W. Aha: The Omnipresence of Case-Based Reasoning in Science and Application. Knowledge-Based Systems, **11**(5-6), 1998

[3] Ch. Alexander, S. Ishikawa und M. Silverstein: A Pattern Language, Oxford University Press, 1977

[4] K.-O. Apel: Der Denkweg von Charles S. Peirce. Suhrkamp Verlag, Frankfurt am Main, 1975

[5] J.C. Arpirez, O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez: WebODE in a nutshell. AI Magazine **24**(3):37-48, 2003

[6] *www.amazon.de*, elektronischer Buch- und Tonträgerhandel

[7] W.R.J. Baets: Organizational Learning and Knowledge Technologies in a Dynamic Environment. Kluwer Academic Publishers 1998

[8] D. Bennet und A. Bennet: The Rise of the Knowledge Organization, in Clyde W. Holsapple: Handbook on Knowledge Management 1, Springer Verlag, Berlin/Heidelberg, 2003

[9] T. Berners-Lee, James Hendler und Ora Lassila: The Semantic Web, Scientific American 2001-05, 2005

[10] G. Bisson, C. Nedellec und L. Canamero (2000), *Designing clustering methods for ontology building - The Mo'K workbench*, Proceedings of the ECAI-2000 workshop on Ontology Learning, Berlin, 2000

[11] P. Bittner, K.-E. Wolff und Ch. Eckes: Conceptual Meaning of Clusters. In Studies in Classification, Data Analysis, and Knowledge Organization, Gaul, W., Locarek-Junge, H. (Herausgeber): Classification in the Information Age; Springer Verlag, 1999

[12] J. Bortz und N. Döring: Forschungsmethoden und Evaluation für Sozialwissenschaftler. 2. Auflage, Springer-Verlag, Berlin, 1995

[13] G. de Chalendar and B. Grau: SVETLAN Or How To Classify Words Using Their Context, in Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2000), Juan-les-Pins, France, 2000

[14] P. Cimiano: Ontology Driven Resolution of Bridging References, In Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5), Tilburg, Niederlande, 2003

[15] R. Coase: The Nature of the Firm, Economica **4**, 1937

[16] D. Cohn und T. Hofmann: The Missing Link: A Probabilistic Model of Document Content and Hypertext Connectivity .In Advances in Neural Information Processing Systems (NIPS **13**), MIT Press, 2001

[17] O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez: Methodologies, tools and languages for building ontologies. Where is the meeting point? Data and Knowledge Engineering **46**(1):41-64, 2003

[18] C. Cortnes und V. Vapnik: Support-Vector Networks, Machine Learning, **20**(3), 1995

[19] C.J. Crouch und B. Yong: Experiments in auto- matic statistical thesaurus construction, SIGIR'92, fünfzehnte

ACM/SIGIR Konerenz: 'Research and Development in Information Retrieval', Kopenhagen, 1992

[20] A. Delteil, C. Faron-Zucker, R. Dieng: Learning Ontologies from RDF annotations, in Proceedings of the Second Workshop on Ontology Learning (OL2001), Seattle, USA, CEUR Workshop Proceedings (CEUR-WS.org), 2001

[21] H. Demmler: Grundlagen der Mikroökonomie. 4. veränderte Auflage, Oldenbourg Verlag, München, 2000

[22] Duden, Das große Wörterbuch der deutschen Sprache in zehn Bänden, Studienausgabe, Duden Verlag, Mannheim, 2003

[23] *www.ebay.com*, elektronische Auktionsplattform

[24] A. El-Saddik: Interactive Multimedia Learning, Springer Verlag, Heidelberg, 2001

[25] A. Faatz und A. Elgert, Concollate, open source collocation detection tool, download and documentation: *http://www.kom.tu-darmstadt.de/Downloads/concollate.html*, Version 0.8.5., 2004

[26] A. Faatz, J. Geise, Achim Steinacker und Ralf Steinmetz: Mappings von Produktkatalogen, Technical Report an der Technischen Universität Darmstadt, 2004

[27] A. Faatz, T. Kamps und R. Steinmetz: Background Knowledge, Indexing and Matching Interdependencies of Document Management and Ontology-Maintenance, Proceedings of the ECAI Worklshop on Ontology Learning, Berlin 2000

[28] C. Fellbaum: A lexical database of English: The mother of all WordNets. Spezialausgabe von 'Computers and the Humanities', Herausgeber: P. Vossen, 1998

[29] R. Fourer, D.M. Gay und Brian W. Kernighan: A Modeling Language for Mathematical Programming, http://www.ampl.com/EXAMPLES/PAPER1/, 1996

[30] J. Fürnkranz, J. Petrak und R. Trappl: Knowledge Discovery in International Conflict Databases. Applied Artificial Intelligence **11**(2):91-118, 1997

[31] A. Gangemi, D. Pisanelli, G. Steve: An overview of the ONIONS project: Applying ontologies to the integration of medical terminologies. Elseviers Journal on Data and Knowledge Engineering **31**, 1999

[32] *http://www.google.de*, die WWW-Suchmaschine Google

[33] N. Guarino und C. Welty: A Formal Ontology of Properties, in Rose Dieng (Herausgeberin), Proceedings der zwölften internationalen Konferenz 'Knowledge Engineering and Knowledge Management', Lecture Notes on Computer Science, Springer Verlag, Heidelberg, 2000

[34] A. Henrich: Management von Softwareprojekten, Oldendourg Verlag, München,2002

[35] E. Gamma, R. Helm, R. Johnson, J. Vlssides: Design Patterns, Elements of Reusable Object-Oriented Software, Addison Wesley Verlag, 1994

[36] A. Gomez-Perez, M. Fernandez-Lopez und O. Corcho: Ontological Engineering, Springer Verlag, 2004

[37] Gregory Grefenstette: Use of Syntactic Context to Produce Term Association Lists for Text Retrieval, SIGIR 1992, Copenhagen, Dänemark, 1992

[38] T. R. Gruber: Towards Principles for the Design of Ontologies Used for Knowledge Sharing. N. Guarino and R. Poli (Eds.): Formal Ontology in Conceptual Analysis and Knowledge Representation,Kluwer Academic Publishers, Deventer, The Netherlands, 1993

[39] *www.hattrick.org*, rechnergestütztes Rollenspiel mit computergenerierter natürlicher Sprache

[40] U. Hahn und C. Schnattinger: Automatic Concept Acquisition From Real-World Texts. In Proceedings of the

AAAI Spring Symposium on 'Machine Learning in Information Access'. AAAI Press, San Mateo, CA, USA, 1998

[41] I.J. Haimowitz, R.S. Patil, P. Szolovits: Representing Medical Knowledge in a Terminological Language is Difficult. in Greenes RA (ed.): Proceedings of the Twelfth SCAMC. Los Angeles, IEEE Computer Society (1988

[42] G. Heyer, U. Quasthoff and Ch. Wolff: Information Extraction from Text Corpora: Using Filters on Collocation Sets, Proceedings der LREC 2002 (Third International Conference On Language Resources And Evaluation), Las Palmas, Spanien, 2002

[43] S. Hörmann, S. Schneider, U. Glowalla, R. Steinmetz: Erstellung von SCORM-kompatiblen Kursen im Projekt k-MED, in Puppe, F.; Albert, J.; Bernauer, J.; Fischer, M.; Klar, R.; Leven, J. (Hrsg.): Proccedingsband des Workshops 'Rechnergestützte Lehr- und Lernsysteme in der Medizin', Universität Würzburg, 2003

[44] T. Hofmann, Lijuan Cai und M. Ciaramita: Learning with Taxonomies: Classifying Documents and Words, Workshop on Syntax, Semantics, and Statistics, Neural Information Processing (NIPS), Vancouver, Kanada 2003

[45] C. W. Holsapple, K. D. Joshi: A collaborative approach to ontology design. Commun. ACM **45**(2): 42-47 ,2002

[46] InformationWeek CMP-WEKA GmbH: SCM - Goldgrube E-Procurement, *www.informationweek.de/print.php3?/channels/ channel08/001066b.htm*, 2000

[47] Institut für medizinische und pharmazeutische Prüfungsfragen Rechtsfähige Anstalt des öffentlichen Rechts: Gegenstandskataloge der Medizin. http://www.impp.de/ImppGk.html, vierte Auflage, 2001

[48] intelligent views GmbH, Softwarewerkzeuge zur Erstellung von Ontologien, www.i-views.de

[49] N. Izumi, T. Yamaguchi: Developing Software Agents Based on Product Ontology and Process Ontology, in Gomez-Perez et. al. (Hrsg.): Tagungsband des Applications of 'Ontologies and Problem-solving Methods'-Workshops, European Conference on Artificial Intelligence 2000, Berlin

[50] T. Joachims: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proceedings of European Conference on Machine Learning(ECML '98), 1998

[51] H. Kim:Predicting how ontologies for the semantic web will evolve, Communications of the ACM, Volume 45 , Issue 2, Februar 2002

[52] k-med: WWW-Ansicht des ontologiebasierten Glossarprototypen http://demo1.intelligent-views.com:3000/kmed/index.html, letzte Aktualisierung 2004

[53] *http://www.k-med.org/*, Projektübersicht zum Projekt k-med(Knowledge based multimedia medical education)

[54] Murray E. Jennex und Lorne Olfman: Organizational Memory, in Clyde W. Holsapple (Hrsg.): Handbook on Knowledge Management 1, Springer Verlag, Berlin/Heidelberg, 2003

[55] Knowledge Interchange Format, draft proposed American National Standard (dpANS), NCITS.T2/98-004, 1998

[56] N.F. Kock, R.J. McQueen und M. Baker: 'Learning and Process Improvement in Knowledge Organisations: A Critical Analysis of Four Contemporary Myths', The Learning Organization **3**, No.1, 1996

[57] T. Kohonen: Self-organizing Maps, dritte Auflage, Springer Verlag, Berlin/Heidelberg, Deutschland, 2001

[58] K. Lagus, S. Kaski, T. Honkela und T. Kohonen: WEBSOM for Textual Data Mining. Artificial Intelligence Review, Volume **13**, 1999

[59] L. J. Lee: Measures of Distributional Similarity , 37th Annual Meeting of the Association for Computational Linguistics, College Park, Maryland, USA, 1999

[60] L. J. Lee: Similarity-Based Approaches to Natural Language Processing. Ph.D. thesis. Harvard University Technical Report TR-11-97, Harvard, USA, 1997

[61] D. B. Lenat und R. V. Guha :Building large knowledge bases. Representation and inference in the Cyc project. Reading et al.: Addison-Wesley Verlag, 1990

[62] K. Lengnink: Formalisierung vom Ähnlichkeit aus Sicht der formalen Begriffsanalyse, Dissertationsschrift, Technische Universität Darmstadt, Shaker Verlag, 1996

[63] Y. Li, Z.A. Bandar and D. McLean: An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources, IEEE Transactions on Knowlege and Data Engineering **15**, 2003

[64] M. Lippert, S. Roock und Henning Wolf: Software entwickeln mit eXtreme Programming - Erfahrungen aus der Praxis, dpunkt.verlag, 2002

[65] A.L. Luft, R. Kötter: Informatik - eine moderne Wissenstechnik. Methodologien der Wissensbildung und Perspektiven der Informatik. BI Wissenschaftsverlag, Mannheim, 1994

[66] H. Lyre: Informationstheorie. Eine philosophisch-naturwissenschaftliche Einführung. Fink/UTB Verlag, München, 2002

[67] A. Mädche: Ontology Learning for the Semantic Web (The Kluwer International Series in Engineering and Computer Science, Volume **665**,2001

[68] A. Mädche, Viktor Pekar, Steffen Staab: Ontology Learning Part One - On Discoverying Taxonomic Relations from the Web, in Ning Zhong et al. (Hrsg.): Web Intelligence. Springer, 2002

[69] D. McGuinness: Ontologies Come of Age, in Dieter Fensel, James Hendler, Henry Liebermann und Wolfgang Wahlster (Hrsg.): Spinning the Semantic Web, MIT Press, Camebridge MA, 2003

[70] H. Maturana, F. Varela: Der Baum der Erkenntnis, aus dem chilenischen Spanisch übersetzt von Kurt Ludewig, Goldmann, München, 1994

[71] T.M. Mitchell: Machine Learning, McGraw Hill, New York, 1997

[72] E. Mittenecker: Planung und statistische Auswertung von Experimenten, Deuticke Verlag, 1971

[73] K. Mudersbach: Begriffe in der Sicht des Sprachnutzers, in Rudolf Wille und Monika Zickwolff (Hrsg.): Begriffliche Wissensverarbeitung, B.I. Wissenschaftsverlag, Mannheim

[74] S. J. Nelson, Tammy Powell und Betsy Humphreys: The Unified Medical Language System (UMLS) Project. In: Kent, Allen; Hall, Carolyn M., editors. Encyclopedia of Library and Information Science. New York: Marcel Dekker, Inc. Verlag, 2002

[75] B. Oestreich: Objektorientierte Softwareentwicklung mit der UML. Oldenbourg, München, fünfte Auflage, 2002

[76] open GALEN reference model, Medizinische Open Source Ontologie, Version 6., 2002 http://www.opengalen.org/open/crm/

[77] E. Ortner: Wissensmanagement, Teil 1: Rekonstruktion des Anwendungswissens. Informatik Spektrum **23**(2): 100-108, 2000

[78] Web Ontology Language (OWL) Use Cases and Requirements, W3C-Empfehlung, herausgegeben von Jeff Heflin, http://www.w3.org/TR/webont-req/, 2004

[79] A. F. Osborn: Applied Imagination, 3. Auflage, Scribner Verlag New York, 1963

[80] F. C. Pereira: Experiments with Free Concept Generation in Divago, in Proceedings of the International Joint Conference on Artificial Intelligence IJCAI'03 Workshop: 3rd Workshop on Creative Systems, Acapulco, 2003

[81] J. Piaget: Biology and Knowledge, Edinburgh University Press, Edinburgh, 1971

[82] M. Plu, P. Bellec, L. Agosto und W. van de Velde: The Web of People, A dual view on the WWW, im Proccedingsband der zwölften WWW-Konferenz, Budapest, 2003

[83] RDF/XML Syntax Specification (revidiert), W3C-Empfehlung, herausgegeben von Dave Beckett, http://www.w3.org/TR/rdf-syntax-grammar/,2004

[84] P. Resnik: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, Journal of Artificial Intelligence Research **11**, 1999

[85] W.A. Russell und O.R. Meseck: Der Einfluß der Assoziation auf das Erinnern von Worten in der deutschen, französischen und englischen Sprache. Zeitschrift für experimentelle und angewandte Psychologie **6**, 191-211, 1959

[86] G. Salton: Automatic term class construction using relevance. A summary of work in automatic pseudoclassification. Information processing and management **16**, 1980

[87] M. Schumacher: Security Engineering with Patterns Origins, Theoretical Models, and New Applications, Lecture Notes in Computer Science , Band **2754**

[88] G. Scragg: Semantic Nets as Memory Models. In E. Charniak und Y. Wilks (Hrsg.): Computational Semantics, Fundamental Studies in Computer Science **4**, north holland, 1976

[89] C. Seeberg: Life Long Learning - Modulare Wissensbasen für elektronische Lernumgebungen. Springer Verlag, Heidelberg, 2002

[90] T. B. Seiler: Begreifen und Verstehen. Ein Buch über Begriffe und Bedeutungen. Darmstädter Schriften zur Allgemeinen Wissenschaft, Band **1**, Verlag Allgemeine Wissenschaft, Darmstadt

[91] J. Sinclair, Corpus Concordance Collocation. Oxford University Press Verlag, 1991

[92] T. Sollazzo, S. Handschuh, S. Staab, M. R. Frank und N. Stojanovic: Semantic Web Service Architecture – Evolving Web Service Standards toward the Semantic Web. FLAIRS Conference, Pensacola, Florida, 2002

[93] J. F. Sowa: Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks/Cole Publishing Co., Pacific Grove, CA, 2000.

[94] J. F. Sowa: Conceptual graphs, draft proposed American National Standard, in W. Tepfenhart und W. Cyre (Hrsg.): Conceptual Structures: Standards and Practices, Lecture Notes in AI **1640**, Springer-Verlag, Berlin, 1999

[95] A. Steinacker: Medienbausteine für web-basierte Lernsysteme, Dissertationsschrift am Fachbereich Informatik der Technischen Universität Darmstadt, Darmstadt 2001

[96] R. Steinmetz und K. Nahrstedt: Multimedia Systems. Springer Verlag, Heidelberg, 2004

[97] G. Stumme: Using Ontologies and Formal Concept Analysis for Organizing Business Knowledge. In: J. Becker,

R. Knackstedt (Eds.) : Wissensmanagement mit Referenzmodellen – Konzepte für die Anwendungssystem- und Organisationsgestaltung, Physica Verlag, Heidelberg 2002

[98] A. Todirascu, L. Romary, D. Bekhouche: Vulcain - An Ontology-Based Information Extraction System, Proceedings of: Natural Language Processing and Information Systems, 6th International Conference on Applications of Natural Language to Information Systems, NLDB 2002, Stockholm, Sweden, June 27-28, 2002

[99] V. Turau: Algorithmische Graphentheorie, Oldenbourg Verlag , München, neueste Auflage 2004

[100] A. Tversky: Features of similarity, Psychological Review **84**, 1977

[101] M. Uschold und R. Jasper: A Framework for Understanding and Classifying Ontology Applications. In KAW99 Twelfth Workshop on Knowledge Acquisition, Modeling and Management. Banff, Alberta, Canada, 1999

[102] J. van Zyl und D. Corbett: Population of a Framework for Understanding and Classifying Ontology Applications. In Proceedings European Conference on Artificial Intelligence, Workshop on Applications of Ontologies and Problem Solving Methods, Berlin, 2000

[103] C. Welty und N. Guarino: Support for Ontological Analysis of Taxonomic Relationships. Journal of Data and Knowledge Engineering. **39**, October, 2001

[104] R. Wille: Concept lattices and conceptual knowledge systems. Computers and Mathematics with Applications **23**, 1992

[105] R. Wille und U. Wille: Restructuring general geometry: measurement and visualization of spatial structures. General Algebra **14**, Johannes Hayn Verlag, Klagenfurt, 2003

[106] G. Witmer: Dictionary of philosophy of mind - ontology. http://www.artsci.wustl.edu/ philos/MindDict/ontology.html, Mai 2000.

[107] L. Wittgenstein: Logisch-Philosophische Abhandlung, erschienen 1921, Neuauflage Suhrkamp Verlag, Frankfurt am Main, 2004

[108] XML Topic Maps (XTM) 1.0, TopicMaps.Org Specification. http://www.topicmaps.org/xtm/1.0/xtm1-20010806.html, 2001

[109] David Yarowsky: Word-Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora, Proceedings der COLING-92, Nantes, Frankreich,1992

Anhang A

Implementierungsdetails

Dieser Anhang zeigt die Umsetzung der ontologischen und vektorwertigen Vergleichsmaße und des Optimierungsproblems für die Testumgebung aus Kapitel 6.

Zentrale Steuerung durch start.pl

Mit der Datei start.pl wird die Anwendung gestartet. Von diesem Skript aus werden die anderen Skripte, die zur Berechnung des Algorithmus notwendig sind, aufgerufen. Der Hauptteil des Quellkodes besteht aus der Implementierung der Benutzeroberfläche.

Formatierung ohne Stopwörter durch conformat.pl

Dieses Skript besitzt die Aufgabe, die in einem Con-Verzeichnis enthaltenen Kollokations-Dateien von unerwünschten Kollokatoren zu befreien. Dies geschieht mit Hilfe einer Stopliste (siehe Anhang). Über die Kommandozeile erhält es einen Zeiger auf das derart zu bearbeitende Con-Verzeichnis, das Verzeichnis, wo das Con-Verzeichnis nach Bearbeitung gespeichert werden soll und einen Zeiger auf die Datei, die die Stopliste enthält. Die Stopliste sollte in einer Textdatei enthalten sein und jede Zeile sollte ein Wort enthalten.
Conformat.pl nimmt zunächst allgemeine Filterungen vor, so werden beispielsweise alle Wörter mit drei oder weniger Buchstaben und alle Wörter, die mit einer Ziffer beginnen, eliminiert. Danach wird die Stopliste Zeile für Zeile mit jeder Kollokations-Datei verglichen. Die übereinstimmenden Einträge werden nicht in die neue Kollokationsdatei übernommen. Folgende Programmzeilen bereinigen die Kollokationsdateien:

```
for ($i = 2; $i < @files; $i++) {#komplettes ConVerzeichnis
durchgehen.
    mkdir ($Argument2);
    #Anlegen des Zielverzeichnisses.
    #Oeffnen der jeweiligen Quell(Con)Datei zur Bearbeitung...

    open (CONDATEI, '$Argument1\\$files[$i]') or die
    ('Datei nicht gefunden');
    #Oeffnen der Zieldatei im Zielverzeichnis, Dateiname wird beibehalten.

    open (CONDATEIFORMATIERT, '>$Argument2\\$files[$i]');
    while (defined($line = <CONDATEI>))
    { #Alle Zeilen der ConDatei durchgehen.

        @zeile = split (',', $line);
        #Array zeile f"ur Zeile der ConDatei.

        if ($zeile[0] =~ /^[^0-9]/)
        { #Alle Eintraege mit Ziffern am Anfang...

         if (length($zeile[0])> 3)
         { #..und alle mit >=3 Zeichen raus.

          $merker = 0;
          #Hilfsvariable merker/Z"ahler (zur"uck-)setzen.

           for ($l = 0; $l < @stopliste; $l++)
           { #Stopliste durchgehen.

            chomp $stopliste[$l];
             if ($stopliste[$l] eq $zeile[0]) {
              $merker = 1;
              #merker/Zaehler setzen bei Uebereinstimmung.
               }#if
                last if ($merker == 1);
                #bei "Ubereinstimmung..

                }#for
                #naechste Zeile untersuchen.
```

```
                if ($merker == 0)  {
                       print CONFORM $line;
                       #wenn keine Uebereinstimmung..

                 }#if
                 #Kollokatoren in neue ConDatei.

               }#if
            }#if
 }#while
    close CONDATEI; #Bearbeitete ConDatei wird
    wieder geschlossen.
```

Wenn der Benutzer Wert darauf legt, dass im Gegensatz zur Standardeinstellung nicht alle Einträge mit drei der weniger Buchstaben entfernt werden, kann in der entsprechenden if-Schleife die getestete Länge herunter gesetzt werden. Das hätte allerdings zur Konsequenz, dass viele dreibuchstabige Worte, wie beispielsweise Präpositionen oder Artikel, die als Kollokatoren keinen Informationsgehalt besitzen, manuell in die Stopliste eingetragen werden müssten.

Vektorkonstruktion

Das Skript Vektorberechnung.pl setzt den ersten Teil des Algorithmus um. Es erhält das Verzeichnis mit den Kollokationsdateien und errechnet daraus die vektorwertigen Vergleichsmaße. Über die Kommandozeile werden die Position des Con-Verzeichnisses und die Angabe, welches vektorwertige Vergleichsmaß verwendet werden soll, übergeben.
Vektorberechnung.pl findet zunächst aus einem gegebenen Con-Verzeichnis jene Kollokatoren, die in mindestens zwei Kollokationsdateien auftreten, und schreibt diese in ein Array. Dieser Vorgang nimmt circa vier Fünftel der Laufzeit des gesamten Programms in Anspruch. Um jeden in einer Kollokationsdatei auftretenden Kollokator mit allen anderen Kollokatoren vergleichen zu können, ist eine vierfach verschachtelte Schleife notwendig.
Der Algorithmus setzt sich aus vier ineinander verschachtelten for-Schleifen zusammen. Die äußerste for-Schleife öffnet nacheinander die zu untersuchenden Kollokationsdateien im Con-Verzeichnis und liest die in der Datei vorhandenen Kollokatoren in ein Array. In der zweiten Schleife werden die Einträge dieses Arrays traversiert. Jeder Eintrag des Arrays wird mit den Kollokatoren aller verbleibenden Kollokationsdateien verglichen. Dies

geschieht jedoch nur dann, wenn der Kollokator nicht schon im Ergebnisarray enthalten ist. Sollte dies der Fall sein, wird sofort zum nächsten Eintrag des Arrays, also zum nächsten Kollokator übergegangen. Wenn der Kollokator noch nicht im Ergebnis-Array enthalten ist, werden nacheinander die anderen Dateien aus dem Con-Verzeichnis geöffnet. Der Laufindex dieser dritten Schleife verhindert dabei, dass ein Kollokator zweimal mit irgendeinem anderen Kollokator verglichen wird. Dies wird durch die Abhängigkeit des Laufindex j der dritten Schleife von dem der äußersten Schleife i mit dem Startwert $j = (i + 1)$ gewährleistet.
Der Inhalt der zu vergleichenden Kollokationsdatei wird wiederum in ein Array geschrieben, das schließlich von der innersten Schleife durchlaufen wird. Innerhalb dieser Schleife wird dann der untersuchte Kollokator mit den Kollokatoren aller anderen Kollokationsdateien verglichen und, wenn er zusätzlich in einer der anderen Datei auftritt, eine Zählvariable gesetzt. Sobald der Zähler für einen Kollokator gesetzt ist, können die beiden inneren Schleifen abgebrochen und der nächste Kollokator untersucht werden. Zuvor wird noch der Kollokator bei gesetztem Zähler an das Ergebnisarray angefügt. Das Ergebnisarray entspricht dem in Kapitel 4 dargestellten Attributmenge $\mathcal{A}$. Der Wortabstand δ_W ist durch die Kollokationsdateien vorgegeben, die Auswahl der zu betrachtenden Kollokatoren, die in $\mathcal{A}$ eingehen, wird durch den soeben beschriebenen Mechanismus getroffen. Diese wird für die spätere Verwendung zur Berechnung der Kandidaten und zur Untersuchung durch den Benutzer, in einer Datei abgespeichert. Das erspart den erneuten Durchlauf der Schleife, wenn in einen folgenden Untersuchungsschritt mit der gleichen Ontologie und dem gleichen Korpus, aber anderen ontologischen und vektorwertigen Vergleichsmaßen angereichert wird.
Nun bildet Vektorberechnung.pl anhand des Vektors für jede Kollokationsdatei weitere Vektoren. Hier werden die Auftrittshäufigkeiten des zur Kollokations-Datei gehörigen Begriffs mit jedem Kollokator des Wortvektors eingetragen. Für jeden Begriff wird dazu dessen Kollokations-Datei geöffnet und die Kollokatoren mit den Einträgen des Wortvektors (der Attributmenge $\mathcal{A}$) verglichen. Bei einer Übereinstimmung wird die Auftrittshäufigkeit dieses Kollokators mit dem Begriff aus der Kollokationsdatei in eine dem Begriff zugehörige Zeile der Matrix @vektor geschrieben. Die sich ergebende Repräsentationsmatrix wird in der Datei Begriffsvektoren.dat zur Verwendung bei der Beurteilung der Kandidaten zwischengespeichert. Durch die Verbindung der Zeilen ergibt sich so die Repräsentationsmatrix für die gegebene Ontologie und für die Attributmenge $\mathcal{A}$. Jeder der Begriffsvektoren wird nun mit sich selbst und allen anderen Vektoren verrechnet. Anhand des übergebenen Kommandozeilenarguments wird entschieden, welches vek-

torwertige Vergleichsmaß zur Erstellung der Matrix verwendet wird. Das matrixförmige Resultat dient dann als unterer Teil der AMPL-Datei, wie wir sie in Abbildung 6.4 vorgestellt haben.
Die Berechnung entspricht den aus den Abschnitten 5.4.2.1 und 5.4.2.2 bekannten Termen für die vektorwertigen Vergleichsmaße.

Berechnung von ontologischen Vergleichsmaßen

Das Skript Ontologieberechnung.pl nimmt die Berechnung der ontologsichen Vergleichsmaße (siehe Abschnitt 5.4.1) und die Formatierung der AMPL-Datei vor. Über die Kommandozeile erhält das Skript Zeiger auf das Con-Verzeichnis, Zeiger auf die Ontologiedatei (wie sie in Abbildung 6.3 dargestellt wurde), Zeiger auf die Datei, in der das Ergebnis zwischengespeichert werden soll, Information über das zu verwendende ontologische Vergleichsmaß (nach Li, Resnik oder dem asymmetrischen Vergleichsmaß), Information darüber, ob Distanz oder Ähnlichkeit berechnet wird und Parameter zur Berechnung des Maßes nach Li.
Des Weiteren geht Ontologieberechnung.pl davon aus, dass sich die Datei Rohdaten.dat mit den berechneten Entfernungsvektoren im Installationsverzeichnis befindet. Zunächst wird die Ontologieadatei eingelesen und die Begriffe der Ontologie werden alphabetisch geordnet. Dies geschieht in der Subroutine ordnen, die über subs.pl in das Skript eingebunden ist.
Bei der Berechnung des asymmetrischen Vergleichsmaßes folgt nun die Bestimmung der Anzahl der in Definition 9 eingeführten Geschwister für jeden Begriff:

```
for $i (0..$#Begriffe) {
        $siblings = 0;
        $siblinghilfvar = 0;
        @erga = split /\./, $Begriffe[$i][0];
        for $j (0..$#Begriffe) {
            @ergb = split /\./, $Begriffe[$j][0];
            if (@erga == @ergb) {
                $a = @erga;
                if ($a == 1) {
                    $siblings++;
                }#if
                else {
```

```
                    for ($c = 0; $c < $a-1 ; $c++) {
                        $siblinghilfvar = 0;
                        last if ($erga[$c] != $ergb[$c]);
                        $siblinghilfvar++;
                    }#for
                #Begriff selbst z"ahlt auch als sibling/Geschwister.
                $siblings = $siblings + $siblinghilfvar;
                }#else
            }#if
        }#for
        $Begriffe[$i][6] = $siblings;
}
```

Ein Begriff ist genau dann ein Geschwister zu einem anderen Begriff, wenn beide Begriffe denselben direkten Oberbegriff besitzen. Daraus folgt, dass die Länge der Positionszahl für beide Begriffe gleich sein muss. Darüber hinaus müssen beide Positionszahlen bis auf die letzte Ziffer gleich sein. Der obige Algorithmus prüft dies und bestimmt dementsprechend die Anzahl der Geschwister. Hier wird der Begriff selbst auch als Geschwister gewertet, was der Definition des asymmetrischen Vergleichsmaßes in Abschnitt 5.4.1.3 entspricht.
Die erhaltenen Anzahl wird für jeden Begriff gespeichert, damit später der Durchschnitt der Anzahl der Geschwister zweier Begriffe berechnet werden kann. Die durchschnittliche Abstraktionsebene wird anhand der Länge der Positionszahl der betrachteten Begriffe bestimmt. Für die Berechnung der Schritte aufwärts und abwärts in der Hierarchie von einem Begriff zum anderen wird folgender Algorithmus verwendet, den wir in der Perl-Notation darstellen:

```
if(@erga < @ergb) {
    $kmax = @erga;
}#if else {
    $kmax = @ergb;
}#else for ($k = 0; $k <= $kmax; $k++) {
    $lambda = @ergb;
    $lambda = $lambda - $k;
    $kappa = @erga;
    $kappa = $kappa - $k ;
    last if ($erga[$k] != $ergb[$k]);
```

```
}
```

Dieser Algorithmus ist sowohl auf die Abwärts- wie auf die Aufwärtsschritte anwendbar. Die Anzahl der Schritte (*kappa* und *lambda* in Gleichung 5.20) hängt von den Positionszahlen der untersuchten Begriffe ab. Je mehr Stellen die Bezeichner von vorne her gemeinsam haben, desto kleiner ist die Anzahl der Schritte innerhalb der Hierarchie. Bei identischen Begriffen ist die Anzahl der Schritte null, bei Begriffen deren Positionszahlen mit unterschiedlichen Ziffern beginnen, entspricht die Anzahl der Schritte der Länge eines der Begriffe, welches Begriffs, ist abhängig von der Richtung, die zur Traversion der Hierarchie verwendet wird.
Der einzige Unterschied zwischen Ähnlichkeit und Unähnlichkeit nach dem asymmetrischen Vergleichsmaß besteht in dem in 5.20 verwendeten Exponenten. Der Variablenname

```
$Distanz
```

wird sowohl für die Distanzberechnung als auch für die Ähnlichkeitsberechnung verwendet. Wurde das Ontologiemaß nach Resnik gewählt, wird nach dem Ordnen der Begriffe der Ontologie zunächst die Subroutine summe aufgerufen. Diese berechnet aus den von Concollate erhaltenen Daten ein Korrelat der absoluten Häufigkeit der Begriffe im Textkorpus, in dem sie die Gesamtauftrittshäufigkeiten aller Kollokatoren für jeden Begriff aufaddiert. Nun wird jeweils die Frequenz berechnet. Die Frequenz eines Begriffs ist hier die Summe der eigenen Gesamtauftrittshäufigkeit im Textkorpus und der aller Begriffe im Unterbaum des Begriffs, das heißt aller direkten und aller weiteren Unterbegriffe bis zu den Begriffen, die keine Unterbegriffe mehr besitzen.

```
for $i (0..$#Begriffe) {
    $Begriffe[$i][4]=0;
    for $j (0..$#Begriffe) {
        #F"ur jeden Begriff wird die Haeufigkeit berechnet.
        if ($Begriffe[$j][0] =~ /^$Begriffe[$i][0]/){
            #Nur der Begriff selbst und seine \\
            #Unterbegriffe werden beruecksichtigt.
            $Begriffe[$i][4] += $Begriffe[$j][2];
        }#if
```

```
    }#for
    $anzahlgesamt += $Begriffe[$i][2];#Bestimmung der Gesamtsumme.
}#for $wertlang = @Werte; #oberster Begriff mit durchschnittlicher
Haeufigkeit wird zusaetzlich eingefuehrt. $oberst =
$anzahlgesamt/$wertlang; $anzahlgesamt = $anzahlgesamt + $oberst;
```

Um Definitionslücken bei der Berechnung des Distanzmaßes zu umgehen wird zusätzlich noch ein bislang nicht explizit aufgeführter abstrakter Oberbegriff $\top$ eingeführt, der als Oberbegriff des bisherigen Wurzelbegriffs der Ontologie zählt und dessen Häufigkeit dem Durchschnitt der Häufigkeiten aller Begriffe entspricht. Nun folgend wird für jeden Begriff die Wahrscheinlichkeit und deren Logarithmus berechnet und im Begriffsarray gespeichert.

```
for $i (0..$#Begriffe) {
    $Begriffe[$i][4] = $Begriffe[$i][4]/$anzahlgesamt;
    #Bei Aehnlichkeit log berechenen...
    if ($ARGV[7] eq '-s' or $ARGV[7] eq '-d') {
    $Begriffe[$i][5]= - log (1-$Begriffe[$i][4]) ;
    }#if
    #Bei Unaehnlichkeit log (1-..) berechnen.
    elsif ($ARGV[7] eq '-j') {
        $Begriffe[$i][5]= - log $Begriffe[$i][4];
    }
}#for
```

Ob Ähnlichkeit oder Unähnlichkeit berechnet wird, ist wiederum von dem gewählten vektorwertigen Vergleichsmaß abhängig (siehe Tabelle 5.3). Nun wird, um die Entfernung zwischen zwei Begriffen zu ermitteln, der Oberbegriff dieser Begriffe ermittelt. Der für diesen Oberbegriff ermittelte Logarithmus entspricht dann der Unähnlichkeit beziehungsweise Ähnlichkeit. Die Ermittlung des Oberbegriffs zweier Begriffe wird im Folgenden gezeigt.

```
for $i (0..$#Begriffe) {
    @erga = split /\./, $Begriffe[$i][0];
    for $j (0..$#Begriffe) {
```

```
        @ergb = split /\./, $Begriffe[$j][0];
        #Oberbegriff berechnen.
        $erg ='';
        if(@erga < @ergb) {
            $zmax = @erga;
        }#if
        else {
            $zmax = @ergb;
        }#else
        if ($erga[0] != $ergb[0]){
            $erg = '0.';
        }#if
        else {
            for($z=0;$z<$zmax;$z++){
                last if($erga[$z]!= $ergb[$z]);
                $erg=$erg.$erga[$z].'.';
            }#for
        }#else
        chop $erg;
    }
}
```

Hier werden nacheinander die Stellen der Positionszahlen der beiden Begriffe verglichen. Für jede Übereinstimmung wird die untersuchte Stelle an den Oberbegriff angefügt. Sobald eine Stelle nicht mehr übereinstimmt wird die Untersuchung abgebrochen und der Oberbegriff ist gefunden. Wenn der Oberbegriff ermittelt ist, wird der zugehörige Vergleichswert in die Ergebnistabelle eingetragen. Bei der Ermittlung der ontologischen Vergleichsmaßes nach Li werden die Abstraktionsebene des Oberbegriffs der beiden untersuchten Begriffe und die Länge des minimalen Pfades zwischen den beiden Begriffen benötigt. Die Ermittlung des Oberbegriffs läuft wie oben beschrieben ab. Es besteht lediglich der Unterschied, dass die Länge der Positionszahl des Oberbegriffs, welche der Abstraktionsebene entspricht, in die Berechnung eingeht. Um die Pfadlänge zu ermitteln, verwenden wie die Algorithmen zur Berechnung der Schritte aufwärts und abwärts in der Hierarchie und addiert diese. Zur Berechnung sind weiterhin die über die Kommandozeile übergebenen Parameter notwendig.

```
#Li Paramter holen. $a = $ARGV[5]; $b = $ARGV[6];


#Unaehnlichkeit berechnen nach Li.

if ($ARGV[7] eq '-s' or $ARGV[7] eq '-d') {

    #Berechnung als Unaehnlichkeitsma"s:
$S4= 1-(exp(-$a*$l) * ((exp($b*$h)-exp(-$b*$h)) / (exp($b*$h) +
exp(-$b*$h)))); } elsif($ARGV[7] eq '-j'){

    #Berechnung als Aehnlichkeitsma"s:
$S4 = exp(-$a*$l) * ((exp($b*$h) - exp(-$b*$h)) / (exp($b*$h) +
exp(-$b*$h))); }


$h
```

entspricht der Abtsraktionsebene des Oberbegriffs und

```
$l
```

der minimalen Pfadlänge zwischen den Begriffen.

Bestimmung von Ähnlichkeitswerten

Nach der Berechnung der ontologischen Vergleichsmaße kann die Beurteilung von möglichen Kandidaten erfolgen. Kandidatenberechnung.pl benötigt hierzu einen Zeiger auf das Verzeichnis mit den Kollokationsdateien der Kandidaten und die Information, welches vektorwertige Vergleichsmaß zur Berechnung verwendet werden soll. Des Weiteren wird dem Skript über die Kommandozeile ein Parameter zur Berechnung des Schiefmaßes übergeben. Zunächst werden die von Vektorberechnung.pl in Dateien zwischengespeicherten Daten, die Attributmenge $\mathcal{A}$ und die Begriffsvektoren, zurückgeholt und in Arrays geschrieben. Ebenso geschieht dies mit der von AMPL errechneten und in einer Datei gespeicherten optimalen Gewichtung.

```
open (VEK, 'Wortvektor.dat'); @ergebnis = <VEK>;

#Wortvektor/Attributmenge zur"uckholen.
```

```
open (BEGVEK, 'Begriffsvektoren.dat'); @Begriffsvektoren=<BEGVEK>;

#Begriffsvektoren/Zeilen der Repr"asentationsmatrix zur"uckholen.

open (KAAS, 'tmp\\loesung.loe'); $line = <KAAS>;

while(defined($line = <KAAS>)) {
    chomp ($line);
    @zeile = split (' ', $line);

    for ($i=0; $i<@zeile; $i = $i+2) {
        $kvek[$zeile[$i]]=$zeile[$i+1];
        #k-Wert holen und nach Gewichtung @kvek schreiben.
    }
}
```

Nach dem Vorbild der Repräsentationsmatrix aus Abschnitt 5.3 wird anhand der Attributmenge $\mathcal{A}$ für jeden Kandidaten ein Kandidatenvektor gebildet. Das bedeutet, dass jeder Kandidat auf sein Vorkommen mit den Kollokatoren aus dem Wortvektor hin überprüft und die Gesamtauftrittshäufigkeit des Kandidaten mit dem Kollokator im Kandidatenvektor gespeichert wird.

```
for ($i = 2; $i < @files; $i++) {
    #Schleife mit 'L"ange des Wortvektors'-Durchl"aufen:
    for ($k=0; $k < @ergebnis; $k++) {
        #ConDatei des Kandidaten wird ge"offnet.
        open (CON, '$Argument0\\$files[$i]')
        or die ('Datei nicht gefunden');
        #ConDatei wird Zeile f"ur Zeile eingelesen
        #und Merker/Zaehler zurueckgesetzt.
        $merka = 0;
        while(defined($line = <CON>)) {
            last if ($merka == 1);
            #Der Inhalt der Zeilen in das array 'zeile'
            #geschrieben.
            @zeile = split(',',$line);
```

```
            $name = $zeile[(@zeile-1)/2];
            chomp $ergebnis[$k];
            #Wenn der Kollokator die Bedingung erf"ullt...
            if ($zeile[0] eq $ergebnis[$k]) {
                $merka = 1;
                #...wird seine Auftrittshaeufigkeit
                #...in das Kandidatenvektorarray geschrieben.
                chomp $zeile[@zeile-1];
                $vektor[$k][$i-2]=$zeile[@zeile-1];
                print SPEICHER $vektor[$k][$i-2],',';
            }#if
        }#while
        if ($merka == 0) {
            $vektor[$k][$i-2]= 0;
            print SPEICHER $vektor[$k][$i-2],',';
        }#if
        #ConDatei wird geschlossen.
        close CON;
    }#for $k
$candidatenames[$i-2] = $name; print SPEICHER $name,'\n'; }#for $i
```

Hierzu wird für jeden Eintrag der Attributmenge $\mathcal{A}$, die sich im Array @ergebnis befindet, jede Kollokations-Datei geöffnet und die darin befindlichen Kollokatoren mit dem jeweiligen Eintrag verglichen. Sobald eine Übereinstimmung in einer Kollokations-Datei gefunden wurde, kann zur nächsten Datei übergegangen werden. Falls ein Eintrag des Wortvektors nicht in einer Kollokations-Datei auftritt wird, an die zugehörige Stelle im Kandidatenvektor eine Null gesetzt, ansonsten die Auftrittshäufigkeit aus der Kollokationsdatei. Nun wird jeder Kandidatenvektor mit jedem Begriffsvektor mit Hilfe des bereits auf das verwendete ontologische Vergleichsmaß abgestimmte vektorwertigen Vergleichsmaßes verrechnet. Nun wird durch Skalarmultiplikation mit der Gewichtung ein reeller Wert berechnet, anhand dessen sich die Eignung des Kandidaten beurteilen lässt.
Im Falle der Verwendung des symmetrischen Jaccardmaßes ist die Ähnlichkeit eines Kandidaten a zu einem Begriff b gleich groß wie die Ähnlichkeit von Begriff b zu Kandidat a. Da das Schiefmaß und das geglättete Schiefmaß jedoch asymmetrische Maße sind, muss sowohl die Distanz von a zu b, als auch die von b zu a berechnet werden. Die berechneten Werte werden schließlich zusammen mit den jeweils verglichenen Kandidaten und Begrif-

fen in eine Datei geschrieben. Anhand der errechneten Entfernungen kann nun beurteilt werden ob sich die Ontologie durch einen der Kandidaten anreichern lässt und an welcher Stelle der Kandidat eingeordnet werden kann.

Anhang B

Weitere Ontologieanwendungen

Der Einsatz mehrerer Ontologien bei einer in der Teilnehmerzahl beschränkten Suchanwendung kann anhand der Abbildung von Produktkatalogen gezeigt werden, ontologiebasierte Spezifikationen anhand der patternbasierten Sicherheit in der Informations- und Kommunikationstechnik. Als Vertreter von Anwendungen mit einer per se unbeschränkten Teilnehmerzahl werden wir den Ansatz des 'Semantic Web' und des 'Web of People' besprechen. Bei beiden verschränken sich die genannten Anwendungsfelder.

Spezifikation

Der folgende Abschnitt beschreibt eine Anwendung von Ontologien auf die strukturierte Spezifikation von Sicherheitslösungen in der Informations- und Kommunikationstechnik. Die Anwendungsklasse, die hier vorgestellt wird, ist somit der Einsatz von Ontologien zur Präzisierung des Sinngehaltes von technischen oder fachbezogenen Spezifikationen.
Die Ausführungen des vorliegenden Abschnittes liefern gleichzeitig ein Beispiel dafür, wie im Bereich der Softwareentwicklung Ontologien eingesetzt werden können. Generell sieht [51] hier das erste zukünftige Anwendungsfeld, das in großem Umfang Ontologien einsetzen wird. Dies sei dadurch bedingt, dass die Anwender in der Softwareentwicklung bereits einen geschulten Zugang zur Formalisierung von Problemen als auch an die Wiederverwendung von technischen Komponenten besitzen. Zudem zieht [51] Parallelen zur Verwendung von Papierformularen, die auch eine Standardisierung des Vokabulars zum Ausfüllen des Formulars nach sich zögen.
Ein Design Pattern ist eine wiederholt anwendbare strukturierte Lösung für

ein konkretes technisches Problem [35]. Der Unterschied zu den im letzten Abschnitt genannten Anwendungen liegt nun zunächst mehr in der Sichtweise als in der Art des Zugriffs auf die Wissensrepräsentation: eine für sich genommenen Spezifikation durch Patterns wird hier bereits bei ihrer Erstellung durch eine vorhandene Ontologie unterstützt. Für die Wiederverwendung von Softwarekomponenten haben [49] Ontologien als Spezifikations- und Suchhilfe herangezogen und eine signifikant niedrigere Einarbeitungs- und Neuentwicklungszeit mit den so dokumentierten Komoponenten festgestellt. Beim Zugriff spielen in dem im Folgenden vorzustellenden Ansatz der Security Patterns zudem noch Schlüsse, die aus der fachgebietsspezifischen Ontologie gezogen werden, eine Rolle.

Der Begriff des Design Patterns stammt ursprünglich aus der Architektur, wo beginnend mit den Arbeiten von [3] wiederkehrende und kombinierbare städtebauliche Lösungen in einer fest vorgegebenen Form tradiert wurden. Die Übertragung dieser Vorgehensweise auf den Bereich des Software Engineering erfolgte ursprünglich parallel mit der Entwicklung der objektorientierten Programmiersprache Smalltalk [35]. Auch hier gibt ein Design Pattern eine wiederholt anwendbare strukturierte Lösung für ein konkretes technisches Problem wider.

Ein Design Pattern im Software Engineering enthält stets die Elemente Name, Kontext, Problem, Randbedingungen und Lösung. Der Name des Design Patterns muss eindeutig und verständlich sein und idealerweise schon viel über den Inhalt des Patterns aussagen. Der Kontext sagt etwas über den Zusammenhang, in welchem Problem und Lösung des Patterns gesehen werden müssen. Randbedingungen hingegen beschreiben Bereiche, die beim Einsatz der Lösung einem positiven oder negativen Nebeneffekt unterliegen.

Die Arbeit von Schumacher [87] benutzt Ontologien für Patterns zum Spezifizieren von Sicherheitslösungen im Software- und Hardwarebereich. Es handelt sich dabei um eine fachgebietsspezifische kleine Ontologie mit zentralen Begriffen aus der Sicherheitstechnik wie 'Bedrohung' (im englischsprachigen Original: threat), 'Angriff' (attack), 'Gegenmaßnahme' (countermeasure). Die Unterbegriffsrelation ist nicht stark ausgeprägt, das heißt, Schumacher definiert in erster Linie auf die fachspezifischen Relationen wie 'verursacht Schaden bei' (causes harm to) oder implementiert (implements). Die Ontologie wurde von Sicherheitsexperten erstellt. Zwischen den im Design Pattern vorhandenen, in natürlicher Sprache ausformulierten Elementen und der Ontologie bestehen dann in der Gesamtspezifikation Relationen wie 'hat Problem' oder 'hat Lösung'. Dies schafft neben einem einheitlichen formalen Rahmen, der bei der Spezifikation benutzt wird und neben den Navigations-, Anfrage- und Suchmöglichkeiten, die im vorhergehenden Abschnitt erläutert

wurden, auch eine formale Möglichkeit der Beurteilung von Sicherheitslösungen. Schumacher untersucht hier so genannte Überdeckungen, mit deren Hilfe festgestellt werden kann, ob bei Anwendung eines Patterns tatsächlich die Wirkungen aller Nebeneffekte der vorgeschlagenen Maßnahmen oder alle zu beseitigenden Ursachen eines Problems erfasst werden.

Abbildung von Produktkatalogen

Der folgende Abschnitt zeigt ein Beispiel, bei dem die Anwendung der Wissensrepräsentation zwei leichtgewichtige Ontologien einsetzt, anstatt von einer weitgehend statischen fachgebietsspezifischen Ontologie auszugehen.
Fast alle größeren Unternehmen nutzen das Internet, um ihre Waren und Dienstleistungen zu verkaufen. Elektronische Markt- und Handelsplätze wie beispielsweise Amazon [6] oder auch ebay [23] zeigen, dass sich das Internet als zusätzlicher Vertriebskanal zum Kunden im so genannten Business-to-Customer-Bereich (B2C) mittlerweile erfolgreich etabliert hat. Der Einsatz von Ontologien in diesem Bereich entspricht den bereits erläuterten Such- und Navigationsanwendungen, wobei hier das Ziel der Suche aus Produkten beziehungsweise besonderen Angeboten des Verkäufers besteht.
Im Gegensatz zum B2C-Bereich wurde die elektronische Beschaffung von Unternehmen selbst bislang weitgehend vernachlässigt: Nur eine Minderheit nutzt derzeit die Möglichkeiten des Internets zur Beschaffung von Gütern und Dienstleistungen. Elektronische Transaktionen im Business-to-Business-Bereich (B2B) finden zumeist mit Hilfe von speziell hierfür konzipierten Lieferantennetzwerken statt, die zum Teil sogar lediglich in Form von 1-zu-1 Verbindungen ohne typische Internetdienste zwischen einem Käufer und einem Lieferanten implementiert sind. Aktuelle Tendenzen zeigen jedoch, dass vor dem Hintergrund eines stetig wachsendem Kostendrucks immer mehr klein- und mittelständische Unternehmen (KMU), Großunternehmen und Konzerne die Kostensenkungspotenziale der elektronisch unterstützen Beschaffung erkennen und versuchen, diese umzusetzen.
Jede Bestellung eines Unternehmens benötigt Ressourcen in Form von Mitarbeitern, unter anderem aus den Bereichen Einkauf, Controlling, Warenannahme, Qualitätskontrolle und Rechnungswesen. Die hierdurch kostenmäßig verursachten Aufwendungen sind im Verhältnis zum Beschaffungsvolumen oft sehr hoch: Laut [46] kalkuliert die Unternehmensberatung KPMG beispielsweise für einen Kugelschreiber im Wert von 2 Euro Prozesskosten in Höhe von bis zu 200 Euro. Durch eine Dezentralisierung des Beschaffungsprozesses - das heißt ausgewählte Einkaufsprozesse können direkt vom Anforderer ausgeführt werden - kann die Einkaufsabteilung von operativen Auf-

gaben entlastet werden. Eine durchgängige datenverarbeitungstechnische Systemintegration der Beschaffungs- und Abrechnungsprozesse ermöglicht die Minimierung von Fehlerquellen und kann so zu einer Kostenersparnis beitragen. Zudem kann eine weit gehende Automatisierung der Bestellabwicklung zur Vereinfachung und Beschleunigung des Beschaffungsvorgangs beitragen. Nicht zuletzt ist eine erhöhte Markttransparenz zu nennen, die durch die Bündelung der Angebote vieler Lieferanten auf einem B2B-Marktplatz im Internet geschaffen werden kann. Diese Transparenz befähigt Unternehmen, Preise und Konditionen für bestimmte Produkte und Dienstleistungen schneller und effizienter miteinander vergleichen zu können als dies im klassischen Bestellprozess mit Hilfe von einzelnen - zum Teil sogar in Papierform vorliegenden - Produktkatalogen möglich ist.
Die zuvor genannten Vorteile verdeutlichen einerseits die Optimierungspotenziale, die sich durch die Nutzung des Informations- und Kommunikationsmediums Internet für die elektronische Unterstützung von Beschaffungsprozessen und deren betriebliche Integration ergeben. Andererseits liefern sie Anhaltspunkte für die bislang fehlende weitflächige Verbreitung von Lösungen: Eine datenverarbeitungstechnische Systemintegration und Automatisierung setzt unter anderem eine syntaktische Standardisierung beispielsweise in Form von Protokollen für den Datenverkehr zwischen den Handelspartnern voraus. Um über virtuelle B2B-Marktplätze im Internet einkaufen oder elektronische Kataloge ins Beschaffungswesen integrieren zu können, ist eine semantische Standardisierung, die eine einheitliche Sprachbasis beziehungsweise Geschäftssprache vereinbart, notwendig: Während es den Beteiligten der traditionellen 1-zu-1 Beziehungen möglich war, sich individuell auf die Terminologie des Partners einzurichten, erfordert ein B2B-Marktplatz - der den Aufbau von Beziehungen zwischen n Lieferanten und m Kunden ermöglicht - ein einheitliches Vokabular, welches durch Ontologien zur Verfügung gestellt werden kann.
Neben der einheitlichen Benennung von Produkten und Dienstleistungen erscheint eine hierarchische Gliederung eben dieser nach einem einheitlichen Schema erforderlich, um den potentiellen Käufern auf einem B2B-Marktplatz eine Funktionalität ähnlich der von zum Beispiel Amazon im B2C-Bereich zur Verfügung stellen zu können. In diesem Zusammenhang spricht man von Klassifikationen oder genauer von Produktklassifikationsstandards. Nach unserer Ontologiedefinition sind hiermit Grundbestandteile einer Ontologie gegeben, wenn wir einen abstrakten Wurzelbegriff $\top$ einführen. Die Produktkataloge sind durch eine Ordnungsrelation hierarchisch gegliedert und die Produkte eindeutig benannt.
Diesen Anforderungen steht, wie die Studie in [46] zeigt, eine ökonomische

Realität entgegen, in der zwar Produktklassifikationsstandards entwickelt werden. Durch ihre Parallelentwicklung jedoch wird der gewünschte Informationsaustausch nicht erleichtert. Da der Einsatz der einzelnen Produktkataloge in den Unternehmen teilweise stark etabliert ist und der Einsatz eines weltweit einheitlichen Klassifikationsstandards einer Idealvorstellung gleichkommt, bleiben nur die Alternativen der Abbildung oder Verschmelzung bestehender Produktklassifikationsstandards.
In [26] werden prototypisch die beiden Produktklassifikationsstandards UNSPSC (Akronym für United Nations Standard Products and Services Code) und eCl@ss aufeinander abgebildet. Die Abbildung erfolgt per Hand und überführt, wenn wir zwei Ontologien nach Definition 4, Ω_1 und Ω_2, in eine gemeinsame Ontologie Ω überführen. Im Gegensatz zur Verschmelzung von Ontologien bleiben die jeweiligen Begriffsmengen B_1 und B_2 erhalten. Für die Menge der Relationen R der Gesamtontologie Ω gilt, wenn wir mit R_1 die Menge der Relationen aus der ersten und mit R_2 die Menge der Relationen aus der zweiten Ontologie bezeichnen:

$$R \supset R_1 \cup R_2 \tag{B.1}$$

Die Relationsmenge R beinhaltet als echte Obermenge von R_1 und R_2 zusätzliche Abbildungsrelationen mit Definitionsbereich B_1 und Wertebereich B_2 und umgekehrt:

$$\forall r \in R \setminus (R_1 \cup R_2) \exists B_r1 \subseteq B_1, B_r2 \subseteq B_2 : B_{r1} = B_1(r), B_{r2} = B_2(r) \tag{B.2}$$

Die zusätzlichen Abbildungsrelationen in R bestehen aus Relationen, die eine Gleichheitsbeziehung oder eine Unterbegriffsbeziehung zwischen den Begriffen aus beiden Ontologien herstellen und aus weiteren semantischen Relationen, die beispielsweise die Funktionalität eines Produktes beschreiben. Die Abbildung B.1 stellt eine Form der Abbildung von UNSPSC und eCl@ss dar.

Die rechte Seite der Abbildung B.1 zeigt einen Ausschnitt der aus eCl@ss gewonnenen Ontologie, links einen Teil aus UNSPSC. Die symmetrische Relation '=' kann als 'ist äquivalent zu' gelesen werden. Innerhalb der in [26] definierten Abbildung von eCl@ss und UNSPSC wird sie für den Fall eingesetzt, dass eine vollständige semantische Gleichheit zwischen den Begriffen (Produktklassen) der beiden Ontologien vorliegt. Auf ähnliche Weise können zusätzliche Unterbegriffsrelationen und eine Relation 'ist ähnlich zu' eingesetzt werden. Wenn beispielsweise Ursache-Wirkungsprinzipien oder Teil-Von-Beziehungen zwischen Begriffen der verschiedenen Ontologien denkbar

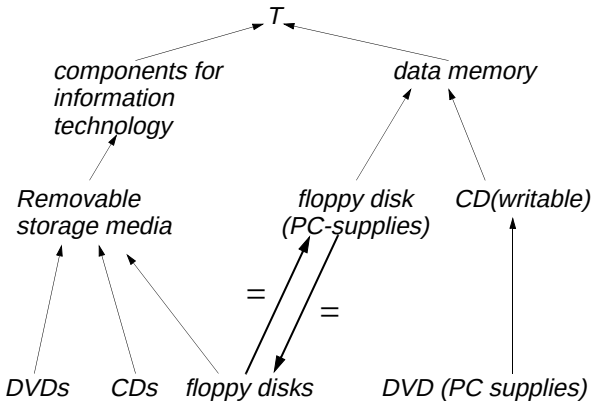

Abbildung B.1: Abbildung zweier Ontologien

sind, so gehören auch diese zur Ausprägung σ der in Gleichung B.1 ausgedrückten echten Obermengenbeziehung.
Die hier mit natürlichsprachlichen Namen versehenen Begriffe sind für jeden Produktklassifikationsstandard mit Ziffernfolgen versehen, die bei einer tatsächlichen Abwicklung eines elektronisch unterstützten Beschaffungsprozesses angegeben werden können. Als Endergebnis wird somit die auf Kundenseite angegebene Kennziffer in die richtige Kennziffer nach dem Klassifikationsstandard des Anbieters überführt.
Die Anwendung von Suche und Navigation in diesem unternehmenübergreifenden Szenario kommt durch die Verbindung zweier oder mehrerer Ontologien zustande. Eine zuverlässige Verbindung muss derzeit per Hand durch einen Experten für Wissensrepräsentation durchgeführt werden. In [26] wurde dazu ein systematisches Vorgehen vorgeschlagen, das unter anderem die Arbeit vom Speziellen zum Allgemeinen und Systematiken beim Namensvergleich von Begriffen vorsieht. Insgesamt stellt sich allerdings eine Kosten-Nutzenfrage, die nur dann positiv beantwortet werden kann, wenn ein vertretbarer Aufwand bei der Abbildung der Ontologien entsteht. Somit sind an dieser Stelle bereits automatisch unterstützte Verfahren in Erwägung zu

ziehen, deren Einsatz den Aufwand bei der manuellen Erstellung der Abbildungsrelationen senken kann.
Die im vorliegenden Abschnitt in Beispielen umrissenen Vorteile und Probleme beim Einsatz mehrerer Ontologien in der Wissensrepräsentation werden auch bei prinzipiell offenen Datenbeständen wie denen des World Wide Web auftreten. Dies zeigt der nächste Abschnitt.

Globale Anwendungen

Wir stellen in diesem Abschnitt zwei Ansätze vor, deren Ziel es ist, die vorausgegangenen Anwendungen von Ontologien für ein sehr großes Szenario einzuführen. Bei de anvisierten Umstrukturierungen des World Wide Web handelt es sich in beiden Fällen um wesentlich umfassendere Ansätze als bei der im Abschnitt 2.3.1 vorgestellten möglichen Verbindung zwischen Ontologien und Suchmaschinen. Vielmehr soll auf einem evolutionären Weg aufbauend auf den bisherigen Ressourcen des WWW eine höhere semantische Interoperabilität geschaffen werden. Bei beiden Ansätzen, der 'Semantic Web'-Initiative des World Wide Web Consortiums W3C und dem 'Web of People' handelt es sich im derzeitigen Stadium noch um Planungen und Visionen, die erst durch eine breite Akzeptanz bei den Anwendern selbst schrittweise verwirklicht werden können.
Das Semantic Web kann als Erweiterung des World Wide Web in einem Schichtenmodell dargestellt werden [92]. Die eigentlichen Inhalte sollen nicht mehr, wie es derzeit der Fall ist, in HTML festgehalten werden, sondern in XML, was eine Trennung von Inhalten und ihrer Darstellung erlaubt. Diese Inhalte sind mit Metadaten versehen, diese bilden die nächste Schicht. Ein Beispiel für Metadaten waren in der Darstellung des Abschnittes 2.3.1 die Schlagwörter für Dokumente. Die für die Metadaten verwendeteten Ausdrücke sind wiederum Begriffe einer von vielen Ontologien. Ein weiteres Merkmal des 'Semantic Web' ist sein dezentraler Aufbau. Die semantische Aufbereitung der Ressourcen im WWW erfolgt nicht über eine gemeinsame große Ontologie, sondern über viele verteilte Ontologien, für deren Zusammenwirken wiederum Abbildungsregeln formuliert sein können. Die weiteren Schichten des 'Semantic Web' oberhalb der Ressourcen, der Metadaten und der Ontologien sind für Logik und Beweisführung (das heißt, für das Festlegen und Anwenden von Schlussregeln bezüglich einer oder mehrerer Ontologien) und für die Bildung von Vertrauenswürdigkeit der vorangegangenen semantischen Schichten zuständig.
Die Anwendungen der Ontologien im 'Semantic Web' unterscheiden sich nicht prinzipiell von den in den vorherigen Abschnitten. Die Gesamtarchi-

tektur zielt jedoch auf einen verstärkten Einsatz von Softwareagenten bei der Nutzung ab. Diese Anwendungen dienen in einem bekannten Beispiel Tim-Berners Lees [9] beispielsweise dazu, Termine zwischen verschiedenen Nutzern des 'Semantic Web' zu koordinieren. An diesem Beispiel wird auch ersichtlich, dass es sich bei den Ontologien im 'Semantic Web' im Gegensatz zu den anderen Beispielen, die wir bislang angeführt haben, nicht nur um fachgebietsspezifische Ontologien handelt, sondern auch um Formalisierungen von Alltagswissen. Die 'Semantic Web' Bewegung ist schließlich auch noch eng mit der Entwicklung von Ontologiesprachen wie OWL (Ontology Web Language [78]) verknüpft, die sich aufgrund ihrer XML-Syntax besonders gut für das WWW eignen. Die verstärkte Nutzung von Softwareagenten setzt eine solche Standardisierung der formalen Sprachen des 'Semantic Web' voraus.

Der Vorschlag zum 'Web of People' [82] stammt von einer Forschungsgruppe der France Telecom. Es handelt sich um eine prototypische Implementierung, während die Entwicklung des 'Semantic Web' teilweise bereits Standardisierungsprozessen unterliegt Hier werden keine neuen Formate bei der Verwaltung von WWW-Ressourcen eingeführt, sondern vor allen Dingen die vorhandenen eindeutigen Lokationen der Dokumente im WWW genutzt. Jeder Teilnehmer des 'Web of People' kann nun selbst eine Ontologie anlegen und zu den Begriffen der Ontologie die eindeutigen Lokationen von Ressourcen sammeln. Für jede solche Sammlung kann nun eine Liste von Personen, die ebenfalls zum 'Web of People' gehören, festgelegt werden. Die Transitivität der Unterbegriffsrelation sorgt nun dafür, dass sowohl für diese mit einem Begriff ausgezeichnete Ressourcenliste als auch für alle Unterbegriffe und die damit verwalteten Ressourcen eine ständige Aktualisierung erfolgt. Die Teilnehmer des 'Web of People' erhalten somit, sofern sie das wünschen, für thematisch ausgezeichnete Bereiche von bestimmten Personen im WWW gefundene, selbst als WWW-Ressource eingestellte oder sogar als Empfänger des 'Web of People' erhaltene Quellen. Durch die Möglichkeit, bestimmte Ressourcen ausschließlich zum eigenen Gebrauch zu verwenden, wird das System zu einer Mischung aus persönlicher Verwaltung und Veröffentlichung oder auch Empfehlung über das WWW. Die gleichzeitige Einführung von persönlichen Ontologien und von an Personen gebundenen Empfehlungen ermöglicht eine deutlichere Vorauswahl. Sie bildet die Grundlage neuartiger Suchmaschinen unter Einbeziehung von Ontologien und der Reputation der Teilnehmer.

Anhang C

Lebenslauf des Verfassers

Ich wurde am 30. November 1972 als Sohn von Friedhelm und Irmgard Antonie Eleonore Faatz (geborene Ehmke) in Friedberg/Hessen geboren. Von 1979 bis 1983 besuchte ich die Grundschule in Bad Nauheim/Schwalheim und von 1983 bis zum Abitur 1992 das St. Lioba Gymnasium in Bad Nauheim. Von 1992 bis 1999 studierte ich an der Justus-Liebig-Universität in Gießen Mathematik und schloss das Studium als Diplom-Mathematiker ab. Von 1999 bis 2004 war ich wissenschaftlicher Mitarbeiter am von Prof. Dr.-Ing. Ralf Steinmetz geleiteten Lehrstuhl für Multimedia Kommunikation (KOM) an der Technischen Universität Darmstadt.

Andreas Faatz